基于局部信息几何的机器学习算法研究

（申请清华大学工学博士学位论文）

培 养 单 位： 电子工程系

学　　科： 信息与通信工程

研 究 生： 徐 祥 祥

指 导 教 师： 张　林　教　授

副指导教师： 黄 绍 伦　助理教授

二〇二〇年七月

A Local Information Geometric Study on Machine Learning Algorithms

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Information and Communication Engineering

by

Xu Xiangxiang

Dissertation Supervisor: Professor Zhang Lin

Associate Supervisor: Assistant Professor Huang Shao-Lun

July, 2020

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘 要

随着大数据时代数据量的快速增长与计算能力的普遍提升，以深度神经网络为代表的机器学习算法广泛应用于图像识别、自然语言处理、推荐系统等应用领域。然而，由于对深度神经网络工作机制认识的欠缺，算法设计及相应的超参数设置往往以经验和启发式的方式进行，在消耗大量计算资源的同时性能无法得到保障。为解决该问题，本文基于信息论中的局部分析方法，对深度神经网络的工作机制及理论特性展开系统性研究，并针对实际应用场景分别设计了有监督学习及无监督学习的高效算法。论文工作的贡献可总结为如下四个部分：

首先，基于对深度神经网络特征提取过程的分析，揭示了该特征提取问题所具有的奇异值分解数学结构。基于该结构，分析表明由神经网络提取的特征也是统计推断中取得最小推断误差的特征，从而为最优特征赋予可解释性。在此基础上，进一步提出了信息论意义上对特征的性能度量，并在实际学习任务上检验了度量的有效性。

其次，基于神经网络特征提取的奇异值分解结构，由奇异向量的微扰分析给出深度神经网络的样本复杂度。本文将样本复杂度刻画为泛化误差所对应的误差指数，分别给出了有监督学习及半监督学习场景下误差指数的解析表达式，并在此基础上考察了训练所需样本数及半监督学习中有标签样本与无标签样本的最优采样策略。

再次，基于对神经网络训练中随机梯度下降法的分析，表明了训练过程中计算效率与平均泛化误差所满足的折中关系，并给出了大样本、小学习率分析机制下平均泛化误差的解析表达式。基于该表达式，本文给出了随机梯度下降法中学习率等重要超参数的理论最优选择，并对实践中具有代表性的超参数调节策略提供了理论解释。

最后，通过将信息论中概率分布的优化问题转化为对特征的优化问题，提出了信息论意义上最优特征求解的深度学习算法设计框架。基于该设计框架，针对有监督分类及无监督多模态数据特征提取的应用场景，分别提出了相应的最优特征学习算法，进一步分析了其与经典机器学习算法的内在联系，并在实际数据集上检验了算法的有效性。

关键词：局部信息几何；机器学习；深度神经网络；样本复杂度；泛化误差

Abstract

With the rapid growth of data amount and the general improvement in the big data era, machine learning algorithms, in particular the deep neural networks (DNNs), have been widely used in applications including image recognition, natural language processing, and recommendation systems. However, due to the lack of theoretical understanding, the design and hyperparameter tuning of DNNs are usually conducted in an empirical and heuristic manner, which consumes much computing resources while provides no performance guarantees. To address this problem, in this thesis, we provide theoretical characterizations of DNNs and design efficient algorithms for supervised learning and unsupervised learning scenarios, respectively, using a local information-theoretic analysis approach. The contribution of this thesis can be summarized in the following four aspects:

Firstly, based on the analysis of feature extraction in DNNs, we reveal the singular value decomposition (SVD) structure of the feature extraction problem. Using this structure, we show that the feature extracted by the network coincides with the one that achieves the minimum inference error in statistical inference tasks. In addition, we propose an information-theoretic performance metric for features and validate this metric in practical learning tasks.

Secondly, using the established SVD nature of feature extraction, we establish the sample complexity of DNNs with a perturbation analysis of singular vectors. In particular, we characterize the sample complexity using the corresponding error exponent of the generalization error and establish the analytical expressions for the exponents for both supervised and semi-supervised learning scenarios. Our results suggest the sample size required for training, together with the optimal sampling strategy for labeled and unlabeled samples in semi-supervised learning.

Thirdly, based on the analysis of the stochastic gradient descent (SGD) for training DNNs, we demonstrate that there exists a tradeoff between computational efficiency and the average generalization error during the training process. In addition, we establish the analytical expression for the average generalization error in the large sample and small learning rate regime. The results suggest the optimal choice for hyperparameters such as the learning rate, and also provide theoretical insights in understanding practical strategies for tuning these parameters.

Finally, we propose a framework to design deep learning algorithms in extracting information-theoretically optimal features, by converting the information-theoretic optimization problem over probability distributions to an optimization problem over features. With this framework, we propose algorithms that extract features for classification tasks and on unsupervised multi-modal data, respectively. We show that the proposed algorithms have deep connections with classical machine learning algorithms and also validate their performances on practical datasets.

Key Words: Local Information Geometry; Machine Learning; Deep Neural Networks; Sample Complexity; Generalization Error

目 录

主要符号对照表

ACE	交替条件期望 (Alternating Conditional Expectation)
SVD	奇异值分解 (Singular Value Decomposition)
X, Y	随机变量
$\mathcal{X}, \mathcal{Y}$	随机变量 X, Y 取值的字母集
$\mathbb{P}(\cdot)$	概率
$\mathbb{E}[\cdot]$	数学期望
$\overset{\mathrm{d}}{=}$	同分布
$\mathcal{P}^{\mathcal{X}}$	所有 $\mathcal{X}$ 上的概率分布的集合
relint$(\cdot)$	集合的相对内部 (Relative Interior)
$\otimes$	矩阵的 Kronecker 乘积
$\circ$	矩阵的 Hadamard 乘积
$\|\cdot\|$	向量的 ℓ_2 范数
$\|\cdot\|_{\mathrm{F}}$	矩阵的 Frobenius 范数
$\|\cdot\|_{\mathrm{s}}$	矩阵谱范数，即最大奇异值
tr$\{\cdot\}$	矩阵的迹 (Trace)
diag$\{d_1, \cdots, d_m\}$	由对角元 $d_1, \cdots, d_m$ 构成的 m 阶对角阵
vec$(\cdot)$	向量化操作。对矩阵 $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{p\times q}$，vec$(\mathbf{W})$ 定义为 pq 维列向量，使其第 $[p(j-1)+i]$ 个元素为 w_{ij}。
$\mathbf{B}$	散度转移矩阵 (Divergence Transfer Matrix，DTM)
$\tilde{\mathbf{B}}$	典型相关矩阵 (Canonical Dependence Matrix，CDM)
$\mathbb{1}_E$	示性函数 (Kronecker 记号)：若事件 E 为真，则 $\mathbb{1}_E = 1$；否则 $\mathbb{1}_E = 0$。
δ_{ij}	Kronecker delta，Kronecker δ 符号：$\delta_{ij} = \mathbb{1}_{i=j}$
$\triangleq$	依定义等于
$\doteq$	渐进相等。给定数列 $\{a_n\}, a_n \doteq \exp(nb)$ 表示 $\lim_{n\to\infty} \frac{1}{n}\log a_n = b$。
η	学习率

第 1 章　绪论

1.1　研究背景

1.1.1　机器学习与神经网络

机器学习是从数据中自动分析获得规律，并利用规律对未知数据进行预测的过程[1]。作为机器学习中的代表性算法，人工神经网络对输入数据进行逐层的特征映射，通过在训练数据上优化映射参数以完成标签预测等任务。图 1.1 给出了一个用于文本识别的神经网络实例。

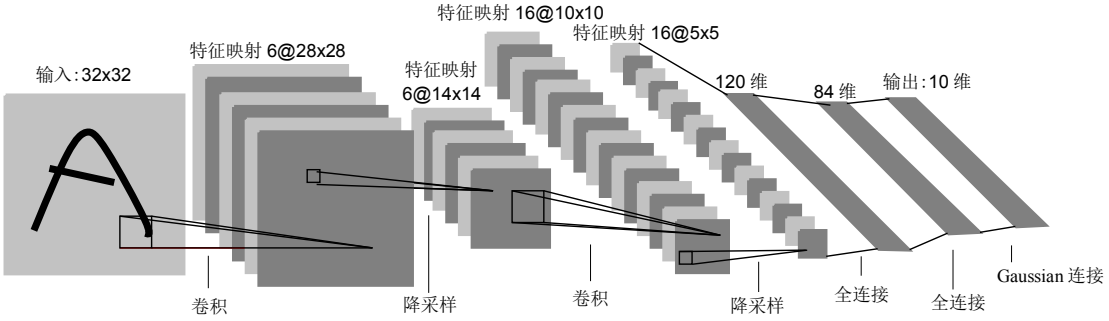


图 1.1　用于文本识别的卷积神经网络 LeNet-5 的结构，共包含七层特征映射 (图源：[2])

得益于大数据时代的海量数据与计算能力的普遍提升，以深度学习为代表的一系列机器学习算法在近些年取得了迅速的发展，掀起了人工智能领域的新一波热潮。通过设计更为复杂的神经网络结构、引入参数优化技巧等方法，深度神经网络表达特征的能力得到极大提升，取得了远胜于经典机器学习方法、甚至是人类的性能，并成功地应用于图像识别、自然语言处理、推荐系统等应用场景[3-4]。在研究领域，新的神经网络结构与设计方法仍持续涌现出来，并在典型数据集上取得不断的性能突破。

然而，对深度学习算法的理论认识仍相当有限，并已成为约束人工智能技术发展的关键瓶颈。具体而言，首先，在算法设计层面，理论基础的欠缺使得深度学习算法的设计及训练多依赖于经验式或启发式的尝试[5-6]，对网络结构参数的微小改动将导致算法的重新训练，由此需要消耗大量的计算能力及时间成本。其次，在算法评价方面，尽管深层网络强大的表达能力可使网络在训练集上取得良好性能，其可能在训练集未出现的潜在数据上表现不佳，即出现过拟合的问题[6]。再次，在算法可靠性层面，由于缺乏性能的理论保障，启发性设计的算法很难用于对安全性、可靠性要求高的应用中，典型的场景包括自动驾驶等。

1.1.2　经典信息论

不同于机器学习构建在数据样本上的实现框架，经典信息论从概率空间角度对信息处理过程给出了建模和分析。自 1948 年 Shannon 奠基性的论文[7] 发表以来，经典信息论在指导通信、信号处理及统计等方面的理论发展及工程实践方面扮演了不可替代的作用[8-9]。具体而言，基于对信源及信道的概率化建模，经典信息论通过对存储、压缩、通信等问题的建模分析，给出了信息处理的数学框架。该框架不仅建立不确定度、信息量等的量化描述，还赋予每个信息度量操作意义 (Operational Meaning)。例如，信息论中的信息熵、交叉熵均对应于信源编码的最优码长，而 K-L 散度 (Kullback-Leibler Divergence, K-L Divergence) 可自然解释为统计学习中对数似然比的期望值、或假设检验问题的误差指数[9]。

虽然经典信息论在理论框架及可解释性方面具有无可比拟的优势，但其目前在除通信领域外的一般数据分析处理方面应用仍十分局限。其中一个重要原因是，信息论的使用依赖于数据概率分布已知的假设。在通信问题中，数据与信道均可建模为相对简单的概率分布，如 Gaussian 分布、Rayleigh 分布等；而在数据分析中，常见的图像、视频等数据很难用简单的概率模型进行描述。以 MNIST 手写体数据集[10] 为例，该数据集中每张图片均为 28×28 的灰度图，每个像素取值为 0 至 255 的整数，如图 1.2 所示。若将数据集中的样本认为是由某一个随机变量生成的独立同分布采样，则该随机变量的可能取值数 (及对应的字母集大小) 为 $256^{28\times28} = 2^{6272} > 10^{1888}$。实践中可供学习的样本数远小于该字母集大小 (如 MNIST 数据集大小为 60,000)，故很难从样本中得出对数据所服从分布的估计，从而信息论的工具难以直接应用于典型的机器学习场景中。

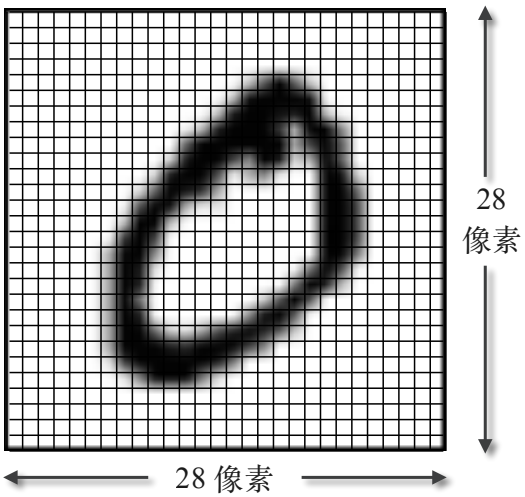


图 1.2　MNIST 手写体数据集中样本示例

1.2　问题引出

针对机器学习算法缺乏理论认识的问题，结合经典信息论的优势，本文对机器学习算法的解释、分析及设计开展研究，具体可分如下四个方面：

1. 为解决深度神经网络的可解释性的问题，基于经典信息论的分析为深度神经网络最优特征赋予操作意义，并由此建立可用于深度神经网络所提取特征的信息度量；
2. 为考察训练集样本数对机器学习算法训练的影响，在所建立信息度量的基础上，考察计算资源充足时算法泛化误差 (Generalization Error) 与样本数的关系，即样本复杂度；
3. 进一步考查机器学习算法训练过程中，实用训练算法 (例如随机梯度下降法) 中超参数选择对训练效率及泛化误差的影响；
4. 最后，借助深度神经网络在特征表达方面的优势，设计基于深度学习框架的应用算法，以实现信息论意义下最优特征的提取。

1.3　研究现状

前述四个问题相关研究可总结如下。

在神经网络可解释性问题上，相关研究对特征及网络参数进行了可视化[11]，以解释神经网络计算机制与传统计算机视觉方法或人类处理图像方式的相似之处。虽然可视化方法可以得到直观的结果，但只能针对单个样本逐个进行，难以获得对特征提取机制的数学理解。除直接可视化特征之外，[12] 利用网络不同层的特征作为输入，分别训练线性分类器，并以分类性能作为对特征好坏的度量，以此解释网络的特征提取机制。此外，Tishby 等将信息瓶颈理论[13-14] 应用于解释神经网络，同时设计了实际网络中的相关实验以验证特征提取与神经网络的关系。尽管这些工作从不同角度、在不同的案例上解释了神经网络所提取特征的优点，但相应结果均依赖于特定的实验，未能建立严格的理论结论。

在机器学习算法样本复杂度问题上，针对特定的问题与学习算法，相关研究给出了泛化误差上界的结果[15]。然而，已有结果高度依赖于具体的问题设定，如对神经网络中激活函数的具体假设[16]、或随机变量之间预先满足的 Markov 关系[17] 等，所得结论难以推广到相类似的问题。此外，由于泛化误差与样本数之间确切关系仍然未知，难以评价相应泛化误差上界 (或样本数的下界) 结论的强弱。

在训练过程超参数选择对泛化误差影响的问题上，针对主成分分析等具体问题，相关研究给出了训练噪声满足特定假设时泛化误差的上界[18-23]。对一般神经网络中的超参数影响的考察多为基于实例的经验性总结，如 [24-25]，而理论工作

局限在如最小二乘等凸问题中[26-27]，对理解实际神经网络性质的帮助有限。

在基于深度神经网络实现信息论意义下最优特征提取方面，由于信息论度量难以从实际数据中精确计算，相关工作通过优化对应度量的上界或者下界、对原优化问题进行松弛、或者引入正则项，以达到求解目的。例如，Belghazi 等通过神经网络实现了对互信息下界的优化，从而可间接用于估计或者优化互信息[28]；又如，Ver Steeg 等讨论了基于信息论中总相关的度量，通过求解松弛后的优化问题以分析数据内在关联的算法[29]。

1.4　研究方法与论文结构安排

本文基于局部信息几何方法对前述问题展开系统性的研究，该方法可有效刻画概率空间中相近的分布之间的关系及由此导出的信息度量。具体而言，基于局部信息几何框架，可将有关概率空间的分布的操作转化为有限维空间对向量的操作，从而概率空间的分析问题可转化为有限维空间的线性代数问题，从而得以有效求解；其典型的应用案例包括统计中假设检验、经典信息论中信道容量求解等问题[30-31]。本文基于局部信息几何方法，建立经典信息论中的信息度量与数据空间在样本上的统计量的联系，使机器学习算法与经典信息论两者的优点得以结合。利用该分析框架，机器学习算法的分析问题均可转化为有限维空间的数学问题，如此既能得到算法理论性质的精确刻画，还可借助经典信息论度量开发高效的新算法。

论文具体结构安排如下：

第 2 章对局部信息几何分析框架进行了简单的介绍。在该框架下，离散随机变量的分布可等价表示为有限维空间中的向量，称为信息向量，并可与随机变量的函数 (特征表示) 建立一一对应关系。在此基础上，经典信息论中的 K-L 散度可表示为信息向量的模长，从而信息论中的一系列运算均可转化为为有限维空间中信息向量的操作。类似地，两个离散随机变量之间的相关性可刻画为相应的矩阵，称为典型相关矩阵，该矩阵范数描述了这两个变量之间的互信息，并可与信息论中经典的 Hirschfeld–Gebelein–Rényi (HGR) 最大相关问题建立紧密联系。

在局部信息几何框架下，第 3 章介绍了对深度神经网络特征提取的理论分析，揭示了深度神经网络特征提取问题与典型相关矩阵的低秩恢复问题的等价性。因此，深度神经网络最优特征可由典型相关矩阵的奇异值分解给出。具体地，神经网络提取特征及网络中对应权重分别对应于典型相关矩阵的左、右奇异向量，也等价于 HGR 最大相关问题中的最优函数，称为最大相关函数。基于该低秩恢复问题结构，进一步给出了神经网络的性能度量，称为 H 评分函数。此外，将局部分

析框架下神经网络特征与权重的对称性推广到了一般情况，并由此建立了神经网络特征与权重的对称关系。

基于第 3 章所建立的神经网络特征提取问题的奇异值分解结构，第 4 章进一步讨论神经网络泛化误差与样本数的依赖关系，即最大相关函数的样本复杂度问题。具体地，借助对矩阵奇异值分解的微扰分析，通过对经验分布空间性质的分析，给出了泛化误差所对应的误差指数的解析表达式。在此基础上，本章考察了半监督学习中有监督样本及无监督样本的比例对误差指数的影响，并由此设计了两类样本采样成本不同时的最优采样机制。

第 5 章考察了实际机器学习应用中广泛采用的参数更新方式，即随机梯度下降法的理论特性。根据第 3 章的结论，局部信息几何框架下的随机梯度下降法可视为求解数据主成分的 Oja 算法[32] 的一个特例。基于此，本章分析了 Oja 算法中计算效率与平均泛化误差的折中关系，考察了随机梯度下降法中超参数选择对泛化误差的影响，并给出最优超参数的选择。利用该折中关系，本章还对残差学习网络结构给出了一种新的理论解释。

此外，借助局部信息几何的分析框架，第 6 章分别介绍了在有监督学习问题与无监督学习问题中，最具信息量特征提取算法的设计，并特别讨论了基于深度学习框架上的算法实现。具体而言，借助局部信息几何的分析框架，可将传统的非线性信息度量转化为信息向量的一系列函数，从而最优的具信息性特征提取问题可转化为信息向量优化的问题，并借助深度学习进行高效求解。在典型数据集上的一系列实验表明，所设计算法与传统方法相比可表现出较明显的性能优势。

最后，第 7 章归纳了以上工作，并对分析方法给出了若干评注。

第 2 章　局部信息几何分析基础

本章对局部信息几何的若干基本概念与结论进行简要介绍，作为后续分析的理论基础。

为方便表述，设 X 为取值于 $\mathcal{X}$ 上的离散随机变量[①]，令 $\mathcal{P}^{\mathcal{X}}$ 表示所有取值在 $\mathcal{X}$ 上随机变量的分布集合，并用 $\mathrm{relint}(\mathcal{P}^{\mathcal{X}})$ 表示该集合相对内部，即 $\mathcal{X}$ 上概率质量函数严格大于 0 的分布集合。

首先引入信息向量的定义，以建立概率分布空间、有限维向量空间 (Euclidean Space) 与函数空间的一一对应关系。

定义 2.1 (信息向量[33])： 给定参考分布 $P_X \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}})$，定义分布 $Q_X \in \mathcal{P}^{\mathcal{X}}$ 所对应的信息向量 $\boldsymbol{\phi}$ 为

$$\phi(x) = \frac{Q_X(x) - P_X(x)}{\sqrt{P_X(x)}} \tag{2-1}$$

并定义函数 $f : \mathcal{X} \mapsto \mathbb{R}$ 使得

$$f(x) = \frac{\phi(x)}{\sqrt{P_X(x)}}, \tag{2-2}$$

由此建立一一对应关系 $Q_X \leftrightarrow \boldsymbol{\phi} \leftrightarrow f$。此外，对于 k 维向量函数 $\mathbf{f} : \mathcal{X} \mapsto \mathbb{R}^k$，令 f_i 表示 $\mathbf{f}$ 的第 i 维，即 $\mathbf{f}(x) = [f_1(x), \cdots, f_k(x)]^{\mathsf{T}}$，定义 $\mathbf{f}$ 等价的矩阵表示 $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{X}| \times k}$ 为

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_k],$$

其中 $\boldsymbol{\phi}_i \leftrightarrow f_i$ 为 f_i 所对应的信息向量表示。

在此基础上，局部信息几何研究概率空间的局部性质，即对给定参考分布小邻域内的概率分布进行分析。该邻域概念可形式化定义如下。

定义 2.2 (ϵ 邻域[34])： 给定参考分布 $P_X \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}})$ 及 $\epsilon > 0$，则 P_X 的 ϵ 邻域定义为

$$\mathcal{N}_{\epsilon}^{\mathcal{X}}(P_X) \triangleq \left\{ Q_X \in \mathcal{P}^{\mathcal{X}} : Q_X \leftrightarrow \boldsymbol{\phi}, \|\boldsymbol{\phi}\| \leqslant \epsilon \right\}.$$

① 因字母集 $\mathcal{X}$ 的元素具体取值不影响分布的属性，为便于表述可不妨假设 $\mathcal{X} = \{1, 2, \cdots, |\mathcal{X}|\}$。

在信息论、统计推断问题及机器学习的分析中，K-L 散度是对不同分布间差异的重要度量。以下命题表明，ϵ 邻域内的分布的 K-L 散度可通过信息向量的长度进行刻画，从而概率空间的分析可转化为对信息向量的分析。

命题 2.1： 给定参考分布 $P_X \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}})$，对所有 $Q_X \in \mathcal{N}_\epsilon^{\mathcal{X}}(P_X)$，$Q_X$ 到 P_X 的 K-L 散度满足[①]

$$D(Q_X \| P_X) \triangleq \sum_{x\in\mathcal{X}} Q_X(x) \log \frac{Q_X(x)}{P_X(x)} = \frac{1}{2}\|\boldsymbol{\phi}\|^2 + o(\epsilon^2)$$

信息向量的概念可自然推广至两个变量的情形。具体地，设 Y 为取值于字母集 $\mathcal{Y}$ 上的离散随机变量，以乘积分布 $P_X P_Y$ 作为参考分布，X 与 Y 的联合分布 P_{XY} 所对应的信息向量可表示为 $|\mathcal{Y}| \times |\mathcal{X}|$ 矩阵，称为*典型相关矩阵*。

定义 2.3： 定义 X 到 Y 的*散度传递矩阵*[35] (Divergence Transfer Matrix，DTM) $\mathbf{B} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ 为

$$B(y,x) \triangleq \frac{P_{XY}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}, \tag{2-3}$$

定义 Y 与 X 的*典型相关矩阵*[33](Canonical Dependence Matrix，CDM) $\tilde{\mathbf{B}} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ 为

$$\tilde{B}(y,x) \triangleq \frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}. \tag{2-4}$$

典型相关矩阵 $\tilde{\mathbf{B}}$ 奇异值分解满足如下性质。

引理 2.1 ([36])： 矩阵 $\tilde{\mathbf{B}}$ 的奇异值分解可表示为 $\tilde{\mathbf{B}} = \sum_{i=1}^{K} \sigma_i \boldsymbol{\psi}_i^Y \left(\boldsymbol{\psi}_i^X\right)^{\mathsf{T}}$，其中 $K \triangleq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$；$\sigma_i$ 表示 $\tilde{\mathbf{B}}$ 的第 i 个奇异值，且满足 $1 \geqslant \sigma_1 \geqslant \cdots \geqslant \sigma_K = 0$；$\boldsymbol{\psi}_i^Y$ 与 $\boldsymbol{\psi}_i^X$ 为对应的左右奇异向量，且 $\psi_K^X(x) = \sqrt{P_X(x)}, \psi_K^Y(y) = \sqrt{P_Y(y)}$。

基于该奇异值分解，可建立 $\tilde{\mathbf{B}}$ 与同样用于刻画随机变量间相关性的广义 HGR 最大相关问题的联系。首先，给出广义 HGR 最大相关的定义如下。

定义 2.4 (广义 HGR 最大相关)： 给定随机变量 X 与 Y 及参数 $k > 0$，其广义 HGR 最大相关定义为

$$\rho_k(X;Y) \triangleq \max_{\substack{\mathbf{f}:\,\mathcal{X}\mapsto\mathbb{R}^k,\,\mathbf{g}:\,\mathcal{Y}\mapsto\mathbb{R}^k \\ \mathbb{E}[\mathbf{f}(X)]=\mathbb{E}[\mathbf{g}(Y)]=\mathbf{0} \\ \mathbb{E}\left[\mathbf{f}(X)\mathbf{f}^{\mathsf{T}}(X)\right]=\mathbb{E}\left[\mathbf{g}(Y)\mathbf{g}^{\mathsf{T}}(Y)\right]=\mathbf{I}}} \mathbb{E}\left[\mathbf{f}^{\mathsf{T}}(X)\mathbf{g}(Y)\right], \tag{2-5}$$

并称取得最优值的 $\mathbf{f}, \mathbf{g}$ 为 k 维 HGR 最大相关函数 (以下简称最大相关函数)。

① 本文中所有对数均为自然对数，即以 e 为底。

特别地，当 $k=1$ 时定义 2.4 为经典的 HGR 最大相关问题[37-39]，对应的 $\rho_1(X;Y)$ 的值称为 HGR 最大相关 (系数)。HGR 最大相关问题是 Pearson 相关系数的重要推广，其可用于作为两个随机变量统计相关性的度量[39]，且广泛应用于机器学习算法设计[33,40-41]。

命题 2.2： k 维最大相关函数 $\mathbf{f}(X)$ 与 $\mathbf{g}(Y)$ 分别对应于[①] $\tilde{\mathbf{B}}$ 的前 k 个右奇异向量及左奇异向量构成的矩阵

$$\begin{bmatrix}\boldsymbol{\psi}_1^X & \cdots & \boldsymbol{\psi}_k^X\end{bmatrix} \quad 及 \quad \begin{bmatrix}\boldsymbol{\psi}_1^Y & \cdots & \boldsymbol{\psi}_k^Y\end{bmatrix}.$$

证明 仅考虑 $k=1$ 的情况，$k>1$ 的情况可类似说明。首先按引理 2.1 的方式定义 σ_i、$\boldsymbol{\psi}_i^X$ 及 $\boldsymbol{\psi}_i^Y$，并设 $f(X)\leftrightarrow\boldsymbol{\phi}^X, g(Y)\leftrightarrow\boldsymbol{\phi}^Y$，则优化问题 (2-5) 可等价表示为

$$\max_{\boldsymbol{\phi}^X,\boldsymbol{\phi}^Y}\left(\boldsymbol{\phi}^Y\right)^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}^X, \tag{2-6}$$

其中 $\boldsymbol{\phi}^X$ 与 $\boldsymbol{\phi}^Y$ 需满足约束条件：$\langle\boldsymbol{\phi}^X,\boldsymbol{\psi}_K^X\rangle=\langle\boldsymbol{\phi}^X,\boldsymbol{\psi}_K^X\rangle=0$ 以及 $\left\|\boldsymbol{\phi}^X\right\|=\left\|\boldsymbol{\phi}^Y\right\|=1$。注意到由于

$$\left(\boldsymbol{\phi}^Y\right)^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}^X\leqslant\left\|\boldsymbol{\phi}^Y\right\|\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{s}}\left\|\boldsymbol{\phi}^X\right\|=\sigma_1\left\|\boldsymbol{\phi}^Y\right\|\left\|\boldsymbol{\phi}^Y\right\|, \qquad \square$$

故优化问题 (2-6) 最大值不超过 σ_1。另一方面，由引理 2.1 知 $\boldsymbol{\phi}^X=\boldsymbol{\psi}_1^X,\boldsymbol{\phi}^Y=\boldsymbol{\psi}_1^Y$ 满足 (2-6) 的约束条件，且对应目标函数值为 σ_1，故其为最优解。

此外，以下性质表明，给定对应关系 $f(X)\leftrightarrow\boldsymbol{\phi}$，典型相关矩阵 $\tilde{\mathbf{B}}$ (或散度转移矩阵) 与 $\boldsymbol{\phi}$ 的乘法操作对应于对特征 $f(X)$ 取条件期望 $\mathbb{E}[\cdot|Y]$ 的操作。

命题 2.3： 对零均值特征 $f(X)$ 及其对应信息向量 $\boldsymbol{\phi}$，有 $\mathbb{E}_{P_{X|Y}}[f(X)|Y]\leftrightarrow\tilde{\mathbf{B}}\boldsymbol{\phi}$。

证明 注意到信息向量 $\tilde{\mathbf{B}}\boldsymbol{\phi}$ 的第 y 个元素为

$$\begin{aligned}\sum_{x\in\mathcal{X}}\tilde{\mathbf{B}}(y,x)\phi(x)&=\sum_{x\in\mathcal{X}}\frac{P_{XY}(x,y)-P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}f(x)\sqrt{P_X(x)}\\&=\frac{1}{\sqrt{P_Y(y)}}\sum_{x\in\mathcal{X}}P_{XY}(x,y)f(x)\\&=\frac{1}{\sqrt{P_Y(y)}}\mathbb{E}_{P_{X|Y}}[f(X)|Y=y].\end{aligned} \qquad \square$$

① 注意这里 $\mathbf{f},\mathbf{g}$ 的选择不唯一：若 $\mathbf{f},\mathbf{g}$ 为最大相关函数，则经过任意 k 阶正交阵 $\mathbf{U}$ 变换后，$\mathbf{Uf}$ 与 $\mathbf{Ug}$ 也为最大相关函数。

第 3 章　深度神经网络的特征提取

3.1　本章引言

针对深度神经网络可解释性不足的问题，本章在局部分析机制下考察神经网络中的特征提取过程。具体地，通过通用特征选择问题的构建，从统计推断角度给出了特征性能的度量，由此导出推断问题中的最优特征。在此基础上，基于局部分析框架，对神经网络 Softmax 层中的特征提取机制进行考察，并建立神经网络所提取的特征与推断问题中最优特征的一致性；该框架可进一步用于神经网络中间隐层的特征提取机制分析。在此基础上，通过考察局部分析框架下神经网络的损失函数，建立了对特征性能的度量，称为 H 评分函数；基于 H 评分函数的对称性，我们进一步研究并证明了神经网络中特征与权重数学上的对称性。

本章具体内容安排如下：第 3.2 节介绍了统计推断中的通用特征选择问题；第 3.3 节考察神经网络 Softmax 层的特征提取机制及其与通用特征选择问题的联系；类似地，第 3.4 节对神经网络隐层的特征提取进行了分析；基于神经网络特征提取的分析结果，第 3.5 节给出了具有操作意义的特征性能度量，第 3.6 节证明了神经网络中特征与权重的对称性；最后，第 3.7 节介绍了神经网络上开展的一系列验证性实验，第 3.8 节对全章内容作了小节。

3.2　通用特征选择问题

给定联合分布为 P_{XY} 的随机变量 X,Y，考察由观测到的 X 的独立同分布样本 $x_1,\cdots,x_n$ 推断 Y 的属性 V 的问题。当统计模型 $P_{X|V}$ 已知时，最优判决准则为对数似然比检验，其中的似然函数可视为推断中的最优特征。但在许多实际问题[36]中，往往难以预先指定目标属性。因此，需转而考虑统计模型未知时具信息性的 X 的低维特征提取问题，称该问题为*通用特征选择问题*。为给出该问题的形式化描述，首先引入相关定义如下。

定义 3.1 (ϵ 相关)： 若 $P_{XY}\in\mathcal{N}_\epsilon^{\mathcal{X}\times\mathcal{Y}}(P_XP_Y)$，则称 X,Y 满足 ϵ 相关。

定义 3.2 (ϵ 属性)： 对给定的 $\epsilon>0$，若随机变量 U 满足对任意 $u\in\mathcal{U}$ 有 $P_{X|U}(\cdot|u)\in\mathcal{N}_\epsilon^{\mathcal{X}}(P_X)$，则称 U 为 X 的 ϵ 属性。

在此基础上，对给定属性 V，称 $\mathcal{C}_{\mathcal{Y}}=\left\{\mathcal{V},\{P_V(v),v\in\mathcal{V}\},\{\boldsymbol{\phi}_v^{Y|V},v\in\mathcal{V}\}\right\}$ 为 V 所对应的*配置*，其中 $\boldsymbol{\phi}_v^{Y|V}\leftrightarrow P_{Y|V}(\cdot|v)$ 为条件分布所对应信息向量。该配置建

模了 V 与 Y 之间的统计相关性。进一步地，这里引入局部分析机制，并假设所考察的 V 均为 Y 的 ϵ 属性，从而其配置满足：对任意 $v \in \mathcal{V}$，有 $\left\|\boldsymbol{\phi}_v^{Y|V}\right\| \leqslant \epsilon$。满足该条件的配置称为$\epsilon$ 配置。此外，假设配置 V 未知，但由某个旋转不变簇 (Rotational Invariant Ensemble，RIE) 生成。

定义 3.3 (旋转不变簇)： 称由

$$\mathcal{C}_{\mathcal{Y}} \triangleq \left\{ \mathcal{V},\ \{P_V(v),\ v \in \mathcal{V}\},\ \{\boldsymbol{\phi}_v^{Y|V},\ v \in \mathcal{V}\} \right\}$$
$$\tilde{\mathcal{C}}_{\mathcal{Y}} \triangleq \left\{ \mathcal{V},\ \{P_V(v),\ v \in \mathcal{V}\},\ \{\tilde{\boldsymbol{\phi}}_v^{Y|V},\ v \in \mathcal{V}\} \right\}.$$

定义的配置 $\mathcal{C}_{\mathcal{Y}}$ 与 $\tilde{\mathcal{C}}_{\mathcal{Y}}$ 旋转等价，若存在正交阵 $\mathbf{Q}$ 使得对任意 $v \in \mathcal{V}$，都有 $\tilde{\boldsymbol{\phi}}_v^{Y|V} = \mathbf{Q}\,\boldsymbol{\phi}_v^{Y|V}$。此外，对定义在一系列配置上的概率测度，若所有旋转等价配置都有相同的测度，则称其为旋转不变簇。

旋转不变簇可解释为是对具有相同可区分度的属性所分配的均匀测度。为了推断属性 V，对给定特征 f_i 及满足 $1 \leqslant k \leqslant K-1$ 的 k，可构造 k 维特征 $h^k = (h_1, \cdots, h_k)$，其中 $h_i = \frac{1}{n}\sum_{l=1}^{n} f_i(x_l),\ i = 1, \cdots, k.$

为求解最优的特征，将考察 f_i 使得基于 h^k 的最优判决准则有最小的错误概率，其中判决性能为由某个旋转不变簇生成的所有可能的配置 $\mathcal{C}_{\mathcal{Y}}$ 上的平均值。为此，令 $\xi_i^X \leftrightarrow f_i$ 表示相应的信息向量，并记 $\boldsymbol{\Xi}^X \triangleq [\xi_1^X\ \cdots\ \xi_k^X]$。

定理 3.1 (通用特征选择)： 给定 $v, v' \in \mathcal{V}$，令 $E_{h^k}(v, v')$ 为根据 h^k 区分 v 及 v' 的误差概率所对应的误差指数，则在由 ϵ 配置所定义的 RIE 上的平均误差指数为

$$\mathbb{E}\left[E_{h^k}(v, v')\right] = \frac{\mathbb{E}\left[\left\|\boldsymbol{\phi}_v^{Y|V} - \boldsymbol{\phi}_{v'}^{Y|V}\right\|^2\right]}{8|\mathcal{Y}|} \left\|\tilde{\mathbf{B}}\boldsymbol{\Xi}^X\left(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X\right)^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2 + o(\epsilon^2), \tag{3-1}$$

其中数学期望操作为 RIE 上不同分布的期望。

证明 参见附录 A.1。 □

由 (3-1) 可知，若将 $\boldsymbol{\xi}_i^X$ 选为 $\tilde{\mathbf{B}}$ 的右奇异向量 $\boldsymbol{\psi}_i^X$ ($i = 1, \cdots, k$)，则 (3-1) 可对所有的 RIE，(v, v') 以及 ϵ 配置取得最优值。因此，$\boldsymbol{\psi}_i^X$ 所对应特征为推断未知属性 V 的 通用最优 (Universally Optimal) 特征。此外，由 (3-1) 可自然导出对 X 的给定特征 $\boldsymbol{\Xi}^X$ 的信息度量 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Xi}^X\left(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X\right)^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2$，其给出了归一化的 $\boldsymbol{\Xi}^X$ 经过线性投影 $\tilde{\mathbf{B}}$ 后的投影长度。该信息度量刻画了 X 的特征在求解关于 Y 的推断问题时具信息性的程度，在特征取 $\tilde{\mathbf{B}}$ 的右奇异向量时达到最优。因此，可将通用特征选择问题解释为求解数据推断问题最具信息性的特征，其对应于 $\tilde{\mathbf{B}}$ 的特征值分解及定义 2.4 中介绍的最大相关函数。接下来将论证，局部分析机制中深度神经网络特征提取的信息度量与通用特征选择中的度量相一致。

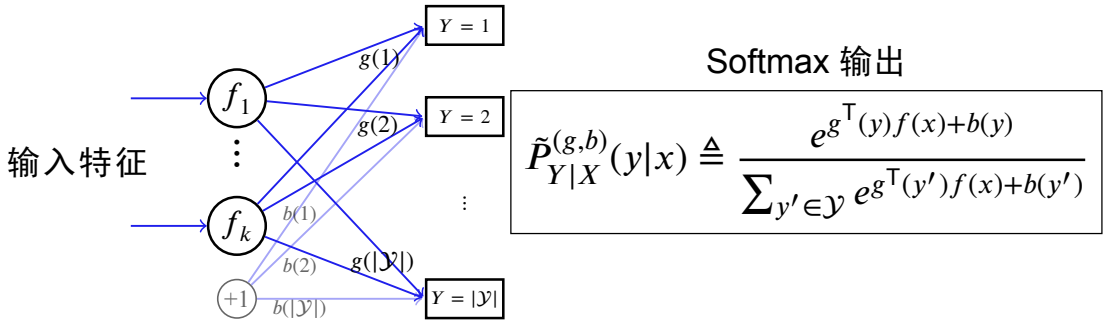


图 3.1　神经网络中的 Softmax 输出层

3.3　Softmax 回归与理想化神经网络

给定数据 X 及对应标签 Y 的样本对 (x_i, y_i)，$i = 1, \cdots, N$，Softmax 回归通过构建形如

$$\tilde{P}^{(g,b)}_{Y|X}(y|x) \triangleq \frac{e^{g^\mathsf{T}(y)f(x)+b(y)}}{\sum_{y'\in\mathcal{Y}} e^{g^\mathsf{T}(y')f(x)+b(y')}} \tag{3-2}$$

的判别模型用于解决分类问题，其中 $f(x) \in \mathbb{R}^k$ 为 X 的 k 维表示，用于预测标签；$g(y) \in \mathbb{R}^k$ 及 $b(y) \in \mathbb{R}$ 的值由以下优化问题决定：

$$(g, b)^* = \arg\max_{(g,b)} \frac{1}{N} \sum_{i=1}^{N} \log \tilde{P}^{(g,b)}_{Y|X}(y_i|x_i). \tag{3-3}$$

普通 Softmax 回归问题对应 $f(x) = x$ 的特例，如图 3.1 所示。更一般地，$f(x)$ 可以为神经网络隐层的输出，即所提取的用于 Softmax 的回归的 x 的特征。接下来的分析将表明，当 X, Y 满足 ϵ 相关性时，函数 $f(x)$ 及 $g(y)$ 与通用特征选择问题的解一致。

令 P_{XY} 为有标签样本 $(x_i, y_i), i = 1, \cdots, N$ 的联合经验分布，并将相应的边缘分布表示为 P_X, P_Y，则优化问题 (3-3) 中的目标函数可表示为对数似然函数的经验平均，亦即 $\frac{1}{N}\sum_{i=1}^{N} \log \tilde{P}^{(g,b)}_{Y|X}(y_i|x_i) = \mathbb{E}_{P_{XY}}\left[\log \tilde{P}^{(g,b)}_{Y|X}(Y|X)\right]$。从而经验平均最大化的问题等价于 K-L 散度最小化问题

$$(g, b)^* = \arg\min_{(g,b)} D(P_{XY} \| P_X \, \tilde{P}^{(g,b)}_{Y|X}), \tag{3-4}$$

其可解释为在形如 $P_X \, \tilde{P}^{(g,b)}_{Y|X}$ 的分布中求解经验联合分布 P_{XY} 的最佳拟合。为便于叙述结论，推导中将偏置项等价表示为 $d(y) = b(y) - \log P_Y(y)$，$y \in \mathcal{Y}$。在此基础上，局部分析机制下问题 (3-4) 的解所满足的约束有如下显式表达。

引理 3.1： 若 X, Y 为 ϵ 相关的随机变量，则 (3-4) 中最优的 g, d 满足[①]

$$|\tilde{g}^{\mathsf{T}}(y) f(x) + \tilde{d}(y)| = O(\epsilon), \quad \forall\, x \in \mathcal{X},\ y \in \mathcal{Y}. \tag{3-5}$$

证明 参见附录 A.2。 □

因此可将 (3-5) 作为局部分析机制下求解问题 (3-4) 的约束条件。此外，定义零均值向量 $\tilde{f}, \tilde{g}$ 所对应的信息向量 $\xi^X(x) = \sqrt{P_X(x)}\,\tilde{f}(x)$，$\xi^Y(y) = \sqrt{P_Y(y)}\,\tilde{g}(y)$，并参照定义 2.1 定义矩阵

$$\begin{aligned} \boldsymbol{\Xi}^Y &\triangleq \left[\xi^Y(1) \quad \cdots \quad \xi^Y(|\mathcal{Y}|)\right]^{\mathsf{T}}, \\ \boldsymbol{\Xi}^X &\triangleq \left[\xi^X(1) \quad \cdots \quad \xi^X(|\mathcal{X}|)\right]^{\mathsf{T}}. \end{aligned}$$

引理 3.2： 在局部分析机制 (3-5) 下，K-L 散度 (3-4) 可表达为

$$D(P_{XY} \| P_X\, \tilde{P}_{Y|X}^{(g,b)}) = \frac{1}{2}\left\|\tilde{\mathbf{B}} - \boldsymbol{\Xi}^Y \left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2 + \frac{1}{2}\gamma^{(g,b)}(f) + o(\epsilon^2), \tag{3-6}$$

其中 $\gamma^{(g,b)}(f) \triangleq \mathbb{E}_{P_Y}\left[(\mu_f^{\mathsf{T}} \tilde{g}(Y) + \tilde{d}(Y))^2\right]$.

证明 参见附录 A.3。 □

该引理给出了神经网络中特征选择的实质。基于 (3-6)，下面进一步考察神经网络中权重、输入特征或者两者都可从数据中训练的情形，对相关的三个问题展开讨论。

3.3.1　前向特征投影

对于给定的 f，可通过固定 $\boldsymbol{\Xi}^X$ 并优化 (3-6) 求得最优权重如下：

定理 3.2： 对固定的 $\boldsymbol{\Xi}^X$ 及 μ_f，使得 (3-6) 最小化的最优的 $\boldsymbol{\Xi}^{Y*}$ 为

$$\boldsymbol{\Xi}^{Y*} = \tilde{\mathbf{B}}\, \boldsymbol{\Xi}^X \left(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}} \boldsymbol{\Xi}^X\right)^{-1}, \tag{3-7}$$

相应的最优权重 $\tilde{g}^*$ 与偏置项 $\tilde{d}^*$ 为

$$\tilde{g}^*(y) = \mathbb{E}_{P_{X|Y}}\left[\boldsymbol{\Lambda}_{\tilde{f}(X)}^{-1} \tilde{f}(X) \,\middle|\, Y = y\right],\ \tilde{d}^*(y) = -\mu_f^{\mathsf{T}}\, \tilde{g}(Y). \tag{3-8}$$

其中 $\boldsymbol{\Lambda}_{\tilde{f}(X)}$ 表示 $\tilde{f}(X)$ 的协方差矩阵。

① 本章的分析中，采用 “~” 符号表示减去均值后的函数，如 $\tilde{g}(y) \triangleq g(y) - \mathbb{E}[g(Y)], y \in \mathcal{Y}$.

证明 参见附录 A.4。 □

式 (3-7) 可视为从输入特征 $\tilde{f}(x)$ 到计算自 y 的特征 $g(y)$ 的投影，其投影结果为与 $\tilde{f}(x)$ 最相关的特征。该投影过程可由左乘 $\tilde{\mathbf{B}}$ 矩阵给出，称之为 “前向特征投影”。

注释 3.1： 上述推导中，我们假设连续输入 $f(x)$ 与离散变量 X 间的函数关系已知，但应用中的计算并不依赖于该函数关系。实际上，条件期望 (3-8) 可直接从 f 与 Y 的样本中计算。因此，若令 f 表示网络输入，上述关于权重及偏置项的分析可直接应用在连续输入的神经网络中。

3.3.2　反向特征投影

类似地可考虑 “反向特征投影” 问题，即在给定权重与偏置项的情况下，求解具信息性的特征 $f^*(X)$ 以最小化损失函数 (3-6) 。

定理 3.3： 对给定的 $\boldsymbol{\Xi}^Y$ 及 $\tilde{d}$，使得 (3-6) 最小化的最优 $\boldsymbol{\Xi}^{X*}$ 为

$$\boldsymbol{\Xi}^{X*} = \tilde{\mathbf{B}}^{\mathsf{T}} \boldsymbol{\Xi}^Y \left(\left(\boldsymbol{\Xi}^Y \right)^{\mathsf{T}} \boldsymbol{\Xi}^Y \right)^{-1}, \tag{3-9}$$

且最优特征 f^* 可分解为 $\tilde{f}^*$ 及 μ_f^* 的和，其中

$$\begin{aligned} \tilde{f}^*(x) &= \mathbb{E}_{P_{Y|X}} \left[\boldsymbol{\Lambda}_{\tilde{g}(Y)}^{-1} \tilde{g}(Y) \middle| X = x \right], \\ \mu_f^* &= -\boldsymbol{\Lambda}_{\tilde{g}(Y)}^{-1} \mathbb{E}_{P_Y} \left[\tilde{g}(Y) \tilde{d}(Y) \right], \end{aligned} \tag{3-10}$$

且 $\boldsymbol{\Lambda}_{\tilde{g}(Y)}$ 表示 $\tilde{g}(Y)$ 的协方差矩阵。

证明 参见附录 A.4。 □

反向特征投影问题解与前向问题解完全对称。在分析中假设特征 $f(X)$ 可取任意形式的函数，即最优特征 (3-10) 可由神经网络生成。该假设只在神经网络具有足够强表达能力时成立，而一般情况网络所表达的函数形式受限于其自身的结构。后面将进一步讨论表达能力受限的情况，并分析神经网络如何近似由 (3-10) 给出的最优特征。

3.3.3　与通用特征选择的联系

当同时优化 f 与 (g, b) (亦即 $\boldsymbol{\Xi}^X, \boldsymbol{\Xi}^Y, d$) 时, 最优的 $(\boldsymbol{\Xi}^Y, \boldsymbol{\Xi}^X)$ 对应于 $\tilde{\mathbf{B}}$ 的低秩分解，从而最优解与通用特征选择问题的解一致。

定理 3.4： 为最小化 (3-6)，特征、权重及偏置项满足：$\tilde{d}(y)=-\mu_f^{\mathsf{T}}\tilde{g}(y)$，且 $(\boldsymbol{\Xi}^Y,\boldsymbol{\Xi}^X)^*$ 分别对应于 $\tilde{\mathbf{B}}$ 前 k 个左右奇异向量。

证明 参见附录 A.5。 □

从而当 f 及 (g,b) 均可优化时，Softmax 回归所提取的特征为输入 X 与标签 Y 间最相关的特征，同时也是通用特征选择问题中对数据推断最具信息性的特征。

注意到结合前向特征投影与反向特征投影的结果，可得到交替求解 $\tilde{\mathbf{f}}^*$ 与 $\tilde{\mathbf{g}}^*$ 的算法如下：

$$\tilde{\mathbf{g}}(y)\leftarrow\boldsymbol{\Lambda}_{\mathbf{f}}^{-1}\mathbb{E}\left[\tilde{\mathbf{f}}(X)\middle|Y=y\right], \tag{3-11a}$$

$$\tilde{\mathbf{f}}(x)\leftarrow\boldsymbol{\Lambda}_{\mathbf{g}}^{-1}\mathbb{E}\left[\tilde{\mathbf{g}}(Y)|X=x\right]. \tag{3-11b}$$

在计算过程中，交替执行 (3-11a) 与 (3-11b)，则 $\tilde{\mathbf{f}}$ 与 $\tilde{\mathbf{g}}$ 可收敛至最优解。该算法可视为对交替条件期望 (Alternating Conditional Expectation，ACE) 算法[42-43] 的多维推广。

在深度神经网络的学习过程中，反向传播算法同时对 Softmax 层及之前所有层的权重进行更新，其中更新权重的操作可理解为前向特征投影 (3-7)，而更新 Softmax 层之前所有层权重的操作可视作反向特征投影(3-9)。因此，反向传播算法可解释为求解 $\tilde{\mathbf{B}}$ 奇异值分解的幂法[44]，即交替条件期望算法。

3.4　表达能力受限的神经网络

根据之前的讨论，Softmax 回归的性能不仅取决于权重及偏置项 $(g(y),b(y))$，也高度依赖于 $f(x)$ 具信息性的程度。事实上，可证明神经网络的隐层本质上也是在提取具信息性的特征。为方便论述，考虑隐层 k 个神经元的神经网络，且隐层输入 $t=[t_1\ \cdots\ t_m]^{\mathsf{T}}\in\mathbb{R}^m$ 均值为零，并假设 t 可表示为离散随机变量[①] X 的函数，记作 $t(x)$。给定有标签样本 $(t(x_i),y_i)$，下面分析该隐层权重与偏置项的最优值。设隐层激活函数为一般的光滑函数 $\sigma(\cdot)$，则第 z 个隐节点 $f_z(X)$ 为

$$f_z(x)=\sigma\left(w^{\mathsf{T}}(z)t(x)+c(z)\right),\quad z=1,\cdots,k,\ x\in\mathcal{X}, \tag{3-12}$$

其中 $w(z)\in\mathbb{R}^m$ 与 $c(z)\in\mathbb{R}$ 分别表示输入层到隐层的权重与偏置项，如图 3.2 所示。此外，令 $f=[f_1\ \cdots\ f_k]^{\mathsf{T}}$ 表示输入 Softmax 层的特征。

① 与注释 3.1 类似, 该离散假设仅仅为了分析方便，实际对权重与偏置项的计算只需了解 t 的信息，不直接依赖于 X。此外，隐层输入 t 可直接从数据获取或由神经网络中上一层输出获得，并统一建模为“预处理”模块，如图 3.2 所示.

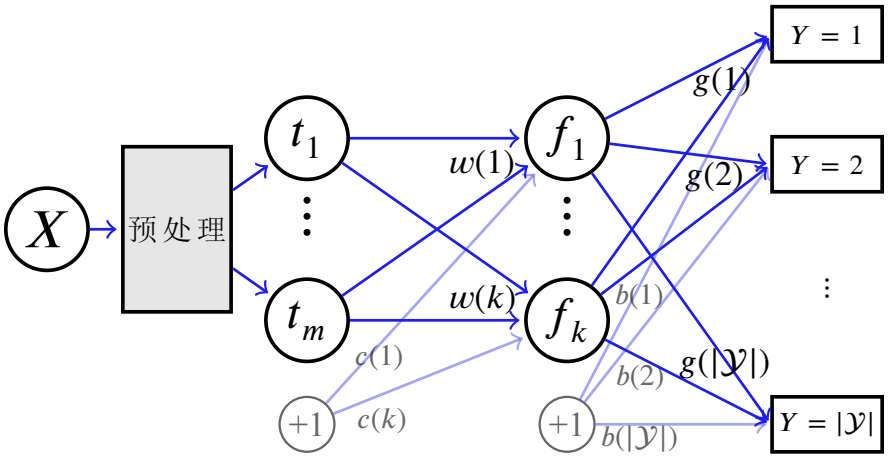


图 3.2　多层神经网络结构，其中 “预处理” 模块包含所有 t 之前的隐层。

为分析隐层的特征提取，固定输出层的 $(g(y), b(y))$，并通过选择最优的 $(w(z), c(z))$ 以最小化输出层 Softmax 回归对应的损失函数 (3-4)。理想情况下 $w(z)$ 及 $c(z)$ 应使得所生成的 $f(x)$ 与(3-10)中最优的 $f^*(x)$ 一致，从而最小化损失函数。但在对 $f(x)$ 的约束条件 (3-12) 下，$f(x)$ 很可能无法取得 $f^*(x)$，而此时网络将通过优化 $w(z), c(z)$ 使得 $f(x)$ 尽量接近 $f^*(x)$。基于局部分析机制，可对该优化过程进行考察如下。

具体地，这里假设相应参数满足局部假设

$$\left|\tilde{g}^{\mathsf{T}}(y) f(x) + \tilde{d}(y)\right| = O(\epsilon),\ \left|w^{\mathsf{T}}(z)\tilde{t}(x)\right| = O(\epsilon),\ \forall x, y, z. \tag{3-13}$$

则由于 t 均值为零，可将 (3-12) 表示为

$$f_z(x) = \sigma\left(w^{\mathsf{T}}(z)t(x) + c(z)\right) = w^{\mathsf{T}}(z)\tilde{t}(x)\cdot\sigma'\left(c(z)\right) + \sigma\left(c(z)\right) + o(\epsilon), \tag{3-14}$$

此外，定义矩阵 $\tilde{\mathbf{B}}_1$ 使其对应元素为 $\tilde{B}_1(z,x) = \frac{\sqrt{P_X(x)}}{\sigma'(c(z))}\tilde{f}_z^*(x)$, 其可视为 CDM 概念在隐层中的推广。进一步地，令 $\xi_1^X(x) = \sqrt{P_X(x)}\,\tilde{t}(x)$ 表示 $\tilde{t}(x)$ 对应的信息向量，并定义矩阵 $\mathbf{\Xi}_1^X \triangleq \left[\xi_1^X(1) \quad \cdots \quad \xi_1^X(|\mathcal{X}|)\right]^{\mathsf{T}}$ 以及

$$\mathbf{W} \triangleq \left[w(1) \quad \cdots \quad w(k)\right]^{\mathsf{T}}$$

$$\mathbf{J} \triangleq \mathrm{diag}\{\sigma'(c(1)), \sigma'(c(2)), \cdots, \sigma'(c(k))\}.$$

损失函数 (3-4) 可由如下定理刻画。

定理 3.5： 给定输出层的权重及偏置项 (g, b)，并将 (g, b) 及 f 所对应的损失函数

(3-4) 简记为 $L(f)$，则在约束 (3-13) 下，有

$$L(f)-L(f^*)=\frac{1}{2}\left\|\Theta\tilde{\mathbf{B}}_1-\Theta\mathbf{W}\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2+\frac{1}{2}\kappa^{(g,b)}(f,f^*)+o(\epsilon^2), \tag{3-15}$$

其中 $\Theta\triangleq\left(\left(\boldsymbol{\Xi}^Y\right)^{\mathsf{T}}\boldsymbol{\Xi}^Y\right)^{1/2}\mathbf{J}$，$\kappa^{(g,b)}(f,f^*)=(\mu_f-\mu_{f^*})^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}(\mu_f-\mu_{f^*})$。

证明 参见附录 A.6。 □

式 (3-15) 给出了从损失函数(3-4)的角度对 f 与 f^* 接近程度的量化。为最小化 (3-15)，可分别考虑如下两个优化问题：

$$\mathbf{W}^*=\arg\min_{\mathbf{W}}\left\|\Theta\tilde{\mathbf{B}}_1-\Theta\mathbf{W}\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2, \tag{3-16}$$

$$\mu_f^*=\arg\min_{\mu_f}\kappa^{(g,b)}(f,f^*). \tag{3-17}$$

首先注意到优化问题 (3-16) 与第 3.3 节中介绍的普通的 Softmax 回归问题类似，且其最优解为 $\mathbf{W}^*=\tilde{\mathbf{B}}_1\boldsymbol{\Xi}_1^X\left(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X\right)^{-1}$。因此，隐层最优权重的求解过程可解释为将 $\tilde{f}^*(x)$ 投影到 $t(x)$ 所张成的函数子空间，从而找到最近的可由神经网络表达的特征。最后，优化问题 (3-17) 的目标是求解 μ_f [亦等价于求解偏置项 $c(z)$] 以最小化与 (3-6) 中 $\gamma^{(v,b)}(s)$ 类似的一个二次项，有关最优解的进一步讨论可参见 (3-17)。

由以上分析可看出 (3-7)，(3-10) 与 (3-16)，(3-17) 之间的对应关系，并将对应操作解释为特征投影。该论断可进一步推广到神经网络中任意一层，具体只需将该层前一层输出视为预处理模块所生成的 $t(x)$，并将其后所有层解释为对 f^* 的优化。从而反向传播算法的计算过程可解释为所有层特征投影的综合。然而即便局部假设成立，由于实际神经网络表达能力有限，最终反向传播算法收敛的解仍可能与 定理 3.4 中奇异值分解的结论不完全一致。该情况下基于特征投影的概念可对网络实际性能与理论性能之间的差距进行度量，该差距同时可作为对所提取特征的度量。

3.5　神经网络性能度量

度量所提取特征对给定学习问题的具信息性程度是机器学习中的基础问题之一[12]。实际上，前述讨论可自然给出对特征有用性的度量。

定义 3.4： 给定特征 $f(x)\in\mathbb{R}^k$ 及权重 $g(y)\in\mathbb{R}^k$，设对应信息矩阵分别为 $\boldsymbol{\Xi}^X$ 和 $\boldsymbol{\Xi}^Y$，定义 H 评分函数[34] $H(f,g)$ 为

$$H(f,g)\triangleq\frac{1}{2}\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2-\frac{1}{2}\left\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2$$

$$= \mathbb{E}_{P_{XY}}\left[\tilde{f}^{\mathsf{T}}(X)\,\tilde{g}(Y)\right] - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Lambda}_{\tilde{f}(X)}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\right). \tag{3-18}$$

此外，定义单边 H 评分函数 $H_Y(f)$ 为

$$H_Y(f) \triangleq \frac{1}{2}\left\|\tilde{\mathbf{B}}\boldsymbol{\Xi}^X\left(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X\right)^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2 = \frac{1}{2}\mathbb{E}_{P_Y}\left[\left\|\mathbb{E}_{P_{X|Y}}\left[\boldsymbol{\Lambda}_{\tilde{f}(X)}^{-1/2}\,\tilde{f}(X)\,\middle|\,Y\right]\right\|^2\right]. \tag{3-19}$$

H 评分函数可用于度量神经网络中间层特征的性能。具体地，可将 $H(f,g)$ 解释为将 $f(x)$ 直接作为 Softmax 输出层的特征，并选取 $g(y)$ 为权重时对应的性能 (损失函数值)。当权重取最优值 $g^*(y)$ 时，相应的性能可用单边 H 评分函数 $H_Y(f)$ 度量，其反映了所提取特征 f 的质量，且与 (3-1) 中导出的信息度量一致。

在机器学习实践中，对数损失函数 (即交叉熵 $\mathbb{E}\left[\log \tilde{P}_{Y|X}^{(g,b)}\right]$) 是目前最常用的性能度量之一，其原则上也可用于评价网络中间层所提取特征的有效性[12]。应用对数损失的一个潜在问题在于，该度量实际意义并不明确，对特定问题其取值甚至可能无界。

与 Log 损失函数不同，H 评分函数可直接从数据样本中计算得出，且由引理 2.1 可知其上界为 $H(f,g) \leqslant H_Y(f) \leqslant \frac{1}{2}\sum_{i=1}^{k}\sigma_i^2 \leqslant k/2$。在这一系列不等式中，第一个 “$\leqslant$” 两边的差刻画了 g 的最优性，第二个 “$\leqslant$” 对应的差量化了所提取特征与最优特征之间区别，其同时揭示了神经网络结构的表达能力；最后的 “$\leqslant$” 对应的差可作为数据集自身好坏的度量。我们将于第 3.7 节中介绍实际数据集中对该度量性质的验证。

3.6　广义 Softmax 学习与神经网络对称性

由 第 3.3 节及 第 3.5 节的讨论可发现，在局部分析机制下，神经网络中 Softmax 层特征 f 与权重 g 在数学上具有对称性。具体而言，可发现 (3-8) 及 (3-10) 形式上具有对称性，定义 3.4 中的 f 与 g 地位上也可互换。另一方面，Softmax 函数 (3-2) 形式上关于 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 是不对称的。为考察局部分析机制下的对称性是否可推广到一般情况，将原 Softmax 回归问题推广为对称的形式，称为*广义 Softmax 学习*[45] (Generalized Softmax Learning, GSL) 问题，定义如下。

定义 3.5： 给定离散随机变量 X, Y 及其联合分布 P_{XY}, 广义 Softmax 学习 (GSL) 问题定义为矩投影 (Moment Projection，M-projection) 问题：

$$\mathcal{M}_{\mathcal{E}_k}(P_{XY}) \triangleq \underset{Q_{XY}\in\mathcal{E}_k}{\arg\min}\, D(P_{XY}\|Q_{XY}), \tag{3-20}$$

其中 $\mathcal{E}_k$ 为形如

$$Q_{XY}(x,y)=\frac{e^{f^{\mathsf{T}}(x)g(y)+a(x)+b(y)}}{\sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}}e^{f^{\mathsf{T}}(x')g(y')+a(x')+b(y')}},\tag{3-21}$$

的指数分布族，且 $f\colon\mathcal{X}\mapsto\mathbb{R}^k$, $g\colon\mathcal{Y}\mapsto\mathbb{R}^k$ 取遍所有可能的 X、Y 的 k 维函数，$a\colon\mathcal{X}\mapsto\mathbb{R}$, $b\colon\mathcal{Y}\mapsto\mathbb{R}$ 为所有可能的标量函数。

下面对 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})$ 的性质及应用展开讨论，其中假设 $P_{XY}\in\operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$。首先注意到指数族 $\mathcal{E}_k$ 可等价表示为

$$\left\{Q_{X,Y}\in\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}\colon Q_{X,Y}(x,y)=e^{f^{\mathsf{T}}(x)g(y)+a(x)+b(y)}\right\}.\tag{3-22}$$

实际上，注意到集合 (3-22) 为 $\mathcal{E}_k$ 的子集。在此基础上，对某个由 (3-21) 给出的 Q_{XY}，定义 $b'(y)\triangleq b(y)-\log\left(\sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}}e^{f^{\mathsf{T}}(x')g(y')+a(x')+b(y')}\right)$，则有 $Q_{XY}(x,y)=e^{f^{\mathsf{T}}(x)g(y)+a(x)+b'(y)}$。在下面的推导中，我们将 $\mathcal{E}_k$ 的分布等价表示为

$$Q_{X,Y}(x,y)=Q[f,g,\alpha,\beta]\triangleq P_X(x)P_Y(y)e^{f^{\mathsf{T}}(x)g(y)-\alpha(x)-\beta(y)},\tag{3-23}$$

其中 $\alpha(x)\triangleq\log P_X(x)-a(x)$，$\beta(y)\triangleq\log P_Y(y)-b(y)$。类似地，使用 $\tilde{P}_{X,Y}=\tilde{P}[f,g,b]$ 表示原 Softmax 回归问题中的联合分布 $\tilde{P}_{X,Y}\triangleq P_X\tilde{P}_{Y|X}$，其中 $\tilde{P}_{Y|X}$ 定义由 (3-2) 给出。此外，这里引入记号 $\mathbf{F}=[f(1),\cdots,f(|\mathcal{X}|)]^{\mathsf{T}}\in\mathbb{R}^{|\mathcal{X}|\times k}$，$\mathbf{G}=[g(1),\cdots,g(|\mathcal{Y}|)]^{\mathsf{T}}\in\mathbb{R}^{|\mathcal{Y}|\times k}$，$\boldsymbol{\alpha}=[\alpha(1),\cdots,\alpha(|\mathcal{X}|)]^{\mathsf{T}}\in\mathbb{R}^{|\mathcal{X}|}$，以及 $\boldsymbol{\beta}=[\beta(1),\cdots,\beta(|\mathcal{Y}|)]^{\mathsf{T}}\in\mathbb{R}^{|\mathcal{Y}|}$。

3.6.1　解的存在性及非唯一性

以下引理可用于解释 (3-20) 解的存在性，其证明可参见附录 A.7。

引理 3.3： 对任意满足 $D(P_{XY}\|Q_{XY})\leqslant D(P_{XY}\|P_XP_Y)$ 的 $Q_{XY}\in\mathcal{E}_k$，存在参数 f,g,α,β 及与 Q_{XY} 独立的常数 $M(P_{XY})$，使得 $Q_{XY}=Q[f,g,\alpha,\beta]$ 且

$$\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\{\|f(x)\|,\|g(y)\|,|\alpha(x)|,|\beta(y)|\}\leqslant M(P_{XY}).\tag{3-24}$$

基于该引理的结果可得如下定理。

定理 3.6： GSL 问题 (3-20) 的解存在。

证明 因 $P_XP_Y=Q[0,0,0,0]\in\mathcal{E}_k$，为求使 $D(P_{XY}\|Q_{XY})$ 最小的 $Q_{XY}\in\mathcal{E}_k$，只需考虑满足 $D(P_{XY}\|Q_{XY})\leqslant D(P_{XY}\|P_XP_Y)$ 的 $Q_{XY}\in\mathcal{E}_k$。根据引理 3.3 可知，满足

该条件的 Q_{XY} 属于集合

$$\begin{aligned}\Big\{Q_{XY}\in\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}:\ &Q_{XY}=Q[f,g,\alpha,\beta],\\ &\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\{\|f(x)\|,\|g(y)\|,|\alpha(x)|,|\beta(y)|\}\leqslant M\Big\},\end{aligned}\tag{3-25}$$

其中 M 为与 Q_{XY} 无关的常数。

从而在紧集 (3-25) 上，关于 Q_{XY} 的连续函数 $D(P_{XY}\|Q_{XY})$ 可取到其最小值。□

一般而言，(3-20) 的解可能不唯一。为说明该性质，首先介绍如下引理，其证明可参见附录 A.8。

引理 3.4： 给定分布 $R_{XY}\in\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$，其逐点互信息 (Pointwise Mutual Information，PMI) 矩阵 $\boldsymbol{\Gamma}=\left[\Gamma_{x,y}\right]\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ 定义为

$$\Gamma_{x,y}\triangleq\log\frac{R_{XY}(x,y)}{R_X(x)R_Y(y)},\quad\forall(x,y)\in\mathcal{X}\times\mathcal{Y}.\tag{3-26}$$

若 $\mathrm{rank}(\boldsymbol{\Gamma})\leqslant k$ 则 $R_{XY}\in\mathcal{E}_k$；且若 $\mathrm{rank}(\boldsymbol{\Gamma})>k+2$，则 $R_{XY}\notin\mathcal{E}_k$。

基于该引理，可构造解不唯一的实例如下。

例 3.1： 考察 GSL 问题，其参数为 $k=1,|\mathcal{X}|=|\mathcal{Y}|=4$，且 P_{XY} 满足

$$P_{XY}(x,y)=u\,\delta_{x,y}+v(1-\delta_{x,y}),\tag{3-27}$$

其中 $u>v>0$，δ 为 Kronecker delta 记号。

设 Q_{XY} 为 $\mathcal{M}_{\mathcal{E}_1}(P_{XY})$ 中的唯一元，则 Q_{XY} 的形式与 P_{XY} 相同，即 $\exists u',v'$ 使得 $Q_{XY}(x,y)=u'\delta_{x,y}+v'(1-\delta_{x,y})$。否则可通过重排 Q_{XY} 中各概率质量的值，得 $Q'_{XY}\in\mathcal{E}_1$ 使其满足 $Q_{XY}\neq Q'_{XY}$ 及 $D(P_{XY}\|Q_{XY})=D(P_{XY}\|Q'_{XY})$，与 Q_{XY} 的唯一性矛盾。

若 $u'\neq v'$，则 Q_{XY} 的 PMI 矩阵满秩，即秩为 $4>k+2=3$，于是由引理 3.4 可知 $Q_{XY}\notin\mathcal{E}_1$。另一方面，由 $u'=v'$ 可推出 $Q_{XY}=P_XP_Y$，此时构造 Q''_{XY} 为

$$Q''_{XY}(x,y)=\begin{cases}u & \text{若 } x=y=1\\ \frac{1-u}{15} & \text{其他情况,}\end{cases}$$

则易知 $Q''_{XY}\in\mathcal{E}_1$ 且 $D(P_{XY}\|Q''_{XY})<D(P_{XY}\|P_XP_Y)$，与矩投影定义矛盾。故 P_{XY} 在 $\mathcal{E}_1$ 上的矩投影不唯一。

3.6.2　几何性质及其应用

本节介绍 GSL 的几何性质及其在机器学习中的应用。

3.6.2.1　GSL 的稳定分布

设 $Q_{XY} = Q[f, g, \alpha, \beta] \in \mathcal{M}_{\mathcal{E}_k}(P_{XY})$，则 (f, g, α, β) 为 GSL 问题 (3-20) 所对应 Lagrange 函数 $\mathcal{L}(f, g, \alpha, \beta, \lambda)$ 的稳定点，其中

$$\mathcal{L}(f, g, \alpha, \beta, \lambda) \triangleq D(P_{XY} \| Q_{XY}) + \lambda \left[\sum_{x', y'} Q_{XY}(x', y') - 1 \right]. \tag{3-28}$$

注意到 $\mathcal{L}$ 的自变量为函数 f, g, α, β 的所有可能取值 (即 $\{f(x), g(y), \alpha(x), \beta(y)\}_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$) 以及 Lagrange 乘子 λ。可验证 $\mathcal{L}$ 的稳定点满足

$$\frac{\partial \mathcal{L}}{\partial f(x)} = \frac{\partial \mathcal{L}}{\partial g(y)} = 0, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \tag{3-29a}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha(x)} = \frac{\partial \mathcal{L}}{\partial \beta(y)} = 0, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \tag{3-29b}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{(x', y') \in \mathcal{X}\times\mathcal{Y}} Q_{XY}(x', y') - 1 = 0. \tag{3-29c}$$

为化简该条件，注意到

$$\begin{aligned} D(P_{XY} \| Q_{XY}) &= \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{Q_{XY}(x, y)} \\ &= D(P_{XY} \| P_X P_Y) - \mathbb{E}_{P_{XY}} \left[f^{\mathsf{T}}(X) g(Y) \right] + \mathbb{E}_{P_X} [\alpha(X)] + \mathbb{E}_{P_Y} [\beta(Y)], \end{aligned} \tag{3-30}$$

从而由 (3-29b) 可知 $P_X = \lambda Q_X, P_Y = \lambda Q_Y$，故 $\lambda = 1$，于是稳定点条件 (3-29) 等价于

$$Q_X = P_X,\ Q_Y = P_Y, \tag{3-31a}$$

$$\mathbb{E}_{Q_{X|Y}} [f(X) \mid Y] = \mathbb{E}_{P_{X|Y}} [f(X) \mid Y], \tag{3-31b}$$

$$\mathbb{E}_{Q_{Y|X}} [g(Y) \mid X] = \mathbb{E}_{P_{Y|X}} [g(Y) \mid X]. \tag{3-31c}$$

称满足 (3-31) 的分布 $Q_{XY} = Q[f, g, \alpha, \beta] \in \mathcal{E}_k$ 为 (GSL 的) 稳定分布 (Stationary Distribution)。令 $\mathcal{E}_k^0$ 表示所有稳定分布的集合，则有 $\mathcal{M}_{\mathcal{E}_k}(P_{XY}) \subset \mathcal{E}_k^0$。此外，由

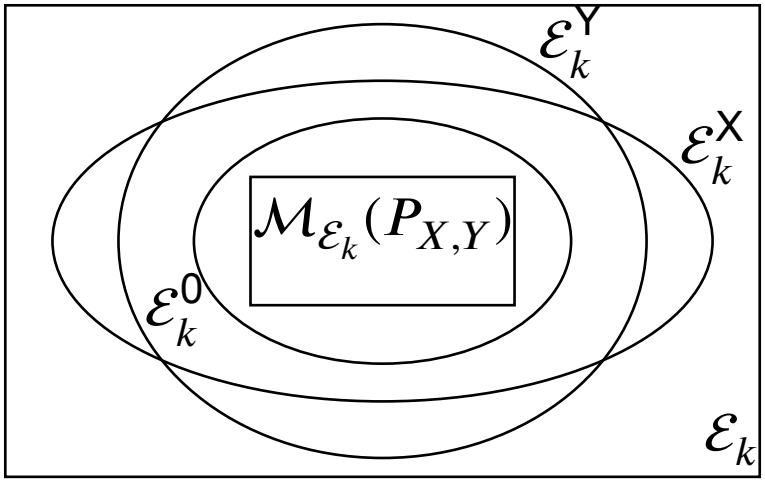


图 3.3　不同分布族间的包含关系

(3-31) 可得 $\mathcal{E}_k^0 \subset \mathcal{E}_k^{\mathsf{X}} \cap \mathcal{E}_k^{\mathsf{Y}}$，其中 $\mathcal{E}_k^{\mathsf{X}}$ 与 $\mathcal{E}_k^{\mathsf{Y}}$ 均为 $\mathcal{E}_k$ 的子集，定义为 $\mathcal{E}_k^{\mathsf{X}} \triangleq \{Q_{XY} \in \mathcal{E}_k : Q_X = P_X\}, \mathcal{E}_k^{\mathsf{Y}} \triangleq \{Q_{XY} \in \mathcal{E}_k : Q_Y = P_Y\}$。这些分布族简单包含关系可参见图 3.3。

集合 $\mathcal{E}_k^0$ 有如下性质。

性质 3.1： $\forall k \in \mathbb{N}_+, \mathcal{E}_k^0 \subset \mathcal{E}_{k+1}^0$.

证明 对任意 $Q_{XY} = Q[f_k, g_k, \alpha, \beta] \in \mathcal{E}_k^0$，可构造 $f_{k+1}(x) = [f_k^{\mathsf{T}}(x), 0]^{\mathsf{T}} \in \mathbb{R}^{k+1}, g_{k+1}(y) = [g_k^{\mathsf{T}}(y), 0]^{\mathsf{T}} \in \mathbb{R}^{k+1}$ 以使得 $Q_{XY} = Q[f_{k+1}, g_{k+1}, \alpha, \beta] \in \mathcal{E}_{k+1}^0$。 □

性质 3.2 (Pythagorean 定理)： $\forall Q_{XY} \in \mathcal{E}_k^0$,

$$D(P_{XY} \| Q_{XY}) + D(Q_{XY} \| P_X P_Y) = D(P_{XY} \| P_X P_Y).$$

证明 对任意 $Q_{XY} = Q[f, g, \alpha, \beta] \in \mathcal{E}_k^0$ 有 $D(Q_{XY} \| P_X P_Y) = \mathbb{E}_{Q_{XY}}\left[f^{\mathsf{T}}(X) g(Y)\right] - \mathbb{E}_{Q_X}[\alpha(X)] - \mathbb{E}_{Q_Y}[\beta(Y)]$。由 $\mathcal{E}_k^0$ 定义 (3-31) 可知 $\mathbb{E}_{Q_X}[\alpha(X)] = \mathbb{E}_{P_X}[\alpha(X)]$ 及 $\mathbb{E}_{Q_Y}[\beta(Y)] = \mathbb{E}_{P_Y}[\beta(Y)]$，从而

$$\begin{aligned}
\mathbb{E}_{Q_{XY}}\left[f^{\mathsf{T}}(X) g(Y)\right] &= \mathbb{E}_{Q_X}\left[f^{\mathsf{T}}(X)\, \mathbb{E}_{Q_{Y|X}}[g(Y)|X]\right] \\
&= \mathbb{E}_{P_X}\left[f^{\mathsf{T}}(X)\, \mathbb{E}_{P_{Y|X}}[g(Y)|X]\right] \\
&= \mathbb{E}_{P_{XY}}\left[f^{\mathsf{T}}(X) g(Y)\right].
\end{aligned}$$

结合以上关系及 (3-30) 可导出如下定理。 □

性质 3.3： $\forall k \in \mathbb{N}_+, \mathcal{M}_{\mathcal{E}_k}(P_{XY}) = \mathcal{E}_k^0$ 当且仅当 $P_{XY} = P_X P_Y$.

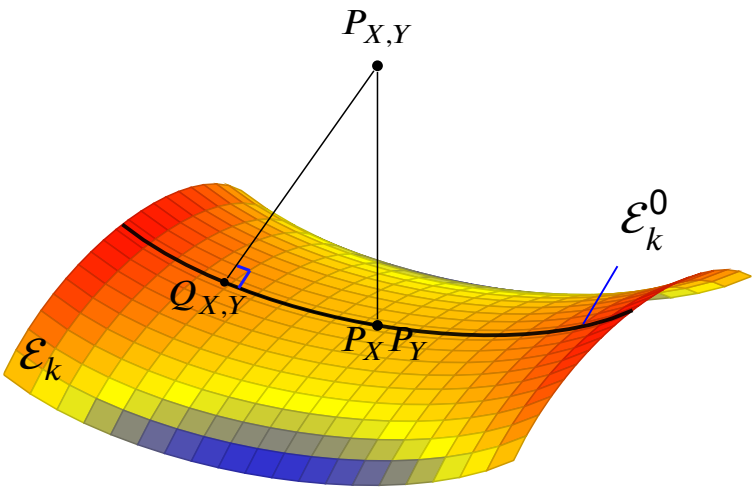


图 3.4　稳定分布 $Q_{XY} \in \mathcal{E}_k^0$ 所满足的 Pythagorean 定理：$D(P_{XY}\|P_XP_Y) = D(P_{XY}\|Q_{XY}) + D(Q_{XY}\|P_XP_Y)$

证明 若 $P_{XY} = P_XP_Y \in \mathcal{E}_k^0$，则根据定义有 $\mathcal{M}_{\mathcal{E}_k}(P_{XY}) = \{P_{XY}\}$。此外，由 性质 3.2 可知，对任意的 $Q_{XY} \in \mathcal{E}_k^0$ 都有 $D(P_{XY}\|Q_{XY}) + D(Q_{XY}\|P_XP_Y) = D(P_{XY}\|P_XP_Y) = 0$，从而 $Q_{XY} = P_{XY} = P_XP_Y$。故 $\mathcal{E}_k^0 = \{P_{XY}\} = \mathcal{M}_{\mathcal{E}_k}(P_{XY})$。

若 $\mathcal{M}_{\mathcal{E}_k}(P_{XY}) = \mathcal{E}_k^0$，则 $P_XP_Y \in \mathcal{M}_{\mathcal{E}_k}(P_{XY})$。因对任意 f 都有 $P_XP_Y = Q[f,0,0,0]$，故 $(f,0,0,0)$ 满足驻点条件 (3-31)。对所有的 $\hat{x} \in \mathcal{X}$，令 $f(x) = [\mathbb{1}_{x=\hat{x}}, 0, \cdots, 0]^\mathsf{T} \in \mathbb{R}^k$，则由 (3-31b) 可得 $\forall y \in \mathcal{Y}, P_{X|Y}(\hat{x}|y) = P_X(\hat{x})$，故有 $P_{XY} = P_XP_Y$。 □

性质 3.1 表明 $\{\mathcal{E}_k^0\}$ 为关于 k 的单调非减集列；性质 3.2 给出了 $\mathcal{E}_k^0$ 上的 Pythagorean 定理，如 图 3.4 所示；性质 3.3 表明，除 X 与 Y 独立的情形外，$\mathcal{M}_{\mathcal{E}_k}(P_{XY})$ 为稳定分布集合 $\mathcal{E}_k^0$ 的真子集。

3.6.2.2　Softmax 学习问题的等价性

为建立 GSL 与原 softmax 回归问题的等价性，这里先给出 Softmax 回归与矩投影问题 $\mathcal{M}_{\mathcal{E}_k^\times}(P_{XY})$ 的等价性。

由 第 3.3 节中讨论可知，Softmax 回归等价于求解

$$\underset{f,g,b}{\text{minimize}}\ D(P_{XY}\|\tilde{P}[f,g,b]), \tag{3-32}$$

其中 P_{XY} 表示数据样本的经验分布，f、g 与 b 分别表示 Softmax 回归中的特征、权重与偏置项。

为表明 (3-32) 与矩投影问题 $\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})$ 的等价性，只需验证 $\mathcal{E}_k^{\mathsf{X}}$ 为形如 $\tilde{P}[f,g,b]$ 的分布族，如以下引理所示。

引理 3.5： 给定参数 f,g，关于分布 R_{XY} 的如下条件等价：

1. $\exists\, b$，使得 $R_{XY}=\tilde{P}[f,g,b]$。
2. $\exists\, \alpha,\beta$，使得 $R_{XY}=Q[f,g,\alpha,\beta]\in\mathcal{E}_k^{\mathsf{X}}$.

证明 1) $\Longrightarrow$ 2)　对给定 f,g,b，设 $\alpha(x)=\log\sum_{y'\in\mathcal{Y}} e^{f^{\mathsf{T}}(x)g(y')+b(y')}$, $\beta(y)=-b(y)+\log P_Y(y)$, 则有 $R_{XY}=\tilde{P}[f,g,b]=Q[f,g,\alpha,\beta]\in\mathcal{E}_k$。又由 $\tilde{P}$ 定义知 $R_X=P_X$，故 $R_{XY}=Q[f,g,\alpha,\beta]\in\mathcal{E}_k^{\mathsf{X}}$。

2) $\Longrightarrow$ 1)　由 $\mathcal{E}_k^{\mathsf{X}}$ 定义知

$$\sum_{y'\in\mathcal{Y}} P_X(x)P_Y(y')e^{f^{\mathsf{T}}(x)g(y')-\alpha(x)-\beta(y')}=P_X(x),$$

从而 $\alpha(x)=\log\sum_{y'\in\mathcal{Y}} P_Y(y')e^{f^{\mathsf{T}}(x)g(y')-\beta(y')}$，故可得 $Q[f,g,\alpha,\beta]=\tilde{P}[f,g,b]$，其中 $b(y)=-\beta(y)+\log P_Y(y)$。 □

原 Softmax 回归问题(3-32)与 GSL 问题 (3-20) 的等价性可由如下定理给出。

定理 3.7： P_{XY} 在集合 $\mathcal{E}_k$、$\mathcal{E}_k^{\mathsf{X}}$ 及 $\mathcal{E}_k^{\mathsf{Y}}$ 上的矩投影相等，亦即

$$\mathcal{M}_{\mathcal{E}_k}(P_{XY})=\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})=\mathcal{M}_{\mathcal{E}_k^{\mathsf{Y}}}(P_{XY}).$$

证明 对任意 $Q_{XY}\in\mathcal{M}_{\mathcal{E}_k}(P_{XY})\subset\mathcal{E}_k^0\subset\mathcal{E}_k^{\mathsf{X}}$，由 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})$ 定义有

$$D(P_{XY}\|Q_{XY})\leqslant D(P_{XY}\|Q'_{XY}),\ \forall\, Q'_{XY}\in\mathcal{E}_k^{\mathsf{X}}\subset\mathcal{E}_k,$$

从而 $Q_{XY}\in\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})$，由此得出 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})\subset\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})$。设 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})\neq\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})$，则对 $Q_{XY}\in\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})\setminus\mathcal{M}_{\mathcal{E}_k}(P_{XY})$，$\exists\, Q'_{XY}\in\mathcal{M}_{\mathcal{E}_k}(P_{XY})\subset\mathcal{E}_k^{\mathsf{X}}$ 使得

$$D(P_{XY}\|Q_{XY})>D(P_{XY}\|Q'_{XY}),$$

从而 $Q_{XY}\notin\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})$，与假设矛盾。故可得 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})=\mathcal{M}_{\mathcal{E}_k^{\mathsf{X}}}(P_{XY})$。同理可得 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})=\mathcal{M}_{\mathcal{E}_k^{\mathsf{Y}}}(P_{XY})$。 □

注释 3.2： 可验证，原 Softmax 回归问题 (3-32) 与 GSL 问题 (3-20) 的稳定解也相同。

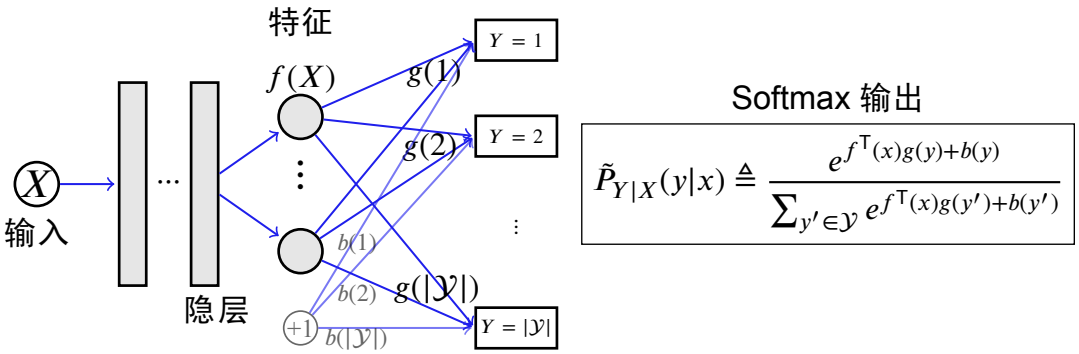


图 3.5　用于分类的前向神经网络，其中 f 为最后的隐层生成的特征，g 及 b 对应最后一层中的权重的偏置项。

3.6.2.3　神经网络中的应用

考察基于数据 X 预测标签 Y 的前向神经网络，如 图 3.5 所示，其中 f 为最后的隐层生成的 X 的特征，g 及 b 分别对应最后一层中的权重与偏置项。当网络中有足够隐神经元时，f 可表示任何函数[46]，因此训练该网络等价于求解原 Softmax 回归问题 (3-32)。以下的分析中，假设神经网络均具有如此的理想表达能力。

由 定理 3.7 可知原 Softmax 回归问题 $\mathcal{M}_{\mathcal{E}_k^{\times}}(P_{XY})$ 及 GSL 问题 $\mathcal{M}_{\mathcal{E}_k}(P_{XY})$ 解集相同。由 f 与 g 在 GSL 问题 (3-20) 中的对称性可知，神经网络提取的特征 f 与权重 g 在训练中的作用也是对称的。

在此基础上，我们有定理 3.7 的直接推论如下，其建立了由 X 预测 Y 的网络及反过来使用 Y 预测 X 的网络间的对称性。

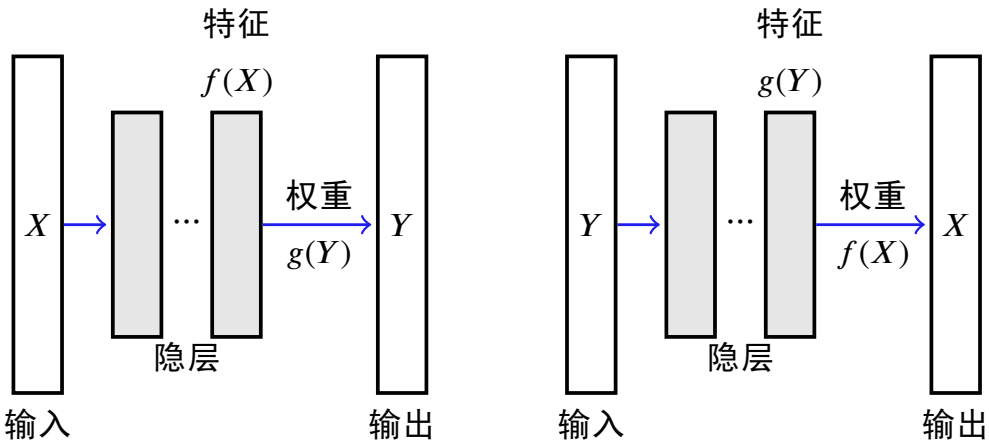
命题 3.1： 基于 X 预测 Y 的神经网络 (X-Y 网络) 与基于 Y 预测 X 的神经网络 (Y-X 网络) 将生成对称的 (特征, 权重) 二元组，如图 3.6 所示。

3.7　实验结果

本节介绍仿真数据集与实际数据集上所开展的一系列实验，以检验前述理论结果。

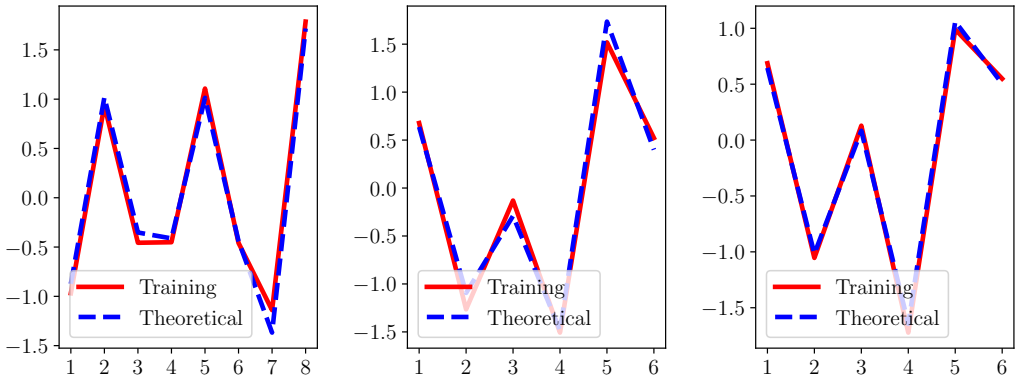
3.7.1　神经网络特征提取

首先检验定理 3.4 中特征投影的结论。为此，取 $k=1$，$|\mathcal{X}|=8$ 以及 $|\mathcal{Y}|=6$，并构造如图 3.5 所示的神经网络，其中网络输入为单点编码后的 X，且利用 Sigmoid 激活的全连接层生成输入特征 $f(X)$。可验证，若全连接层中权重取合适的值，该全连接层可输出任意 X 的函数 (不计平移及等比例放缩变换)，即该神经网络具有理想表达能力。为将网络训练结果与定理 3.4 中理论值对比，首先随机生成分布



(a) X-Y 网络　　(b) Y-X 网络

图 3.6　由 X-Y 网络与 Y-X 网络生成的 (特征, 权重) 二元组的对称关系



(a) $f(x)$　　(b) $g(y)$　　(c) $b(y)$

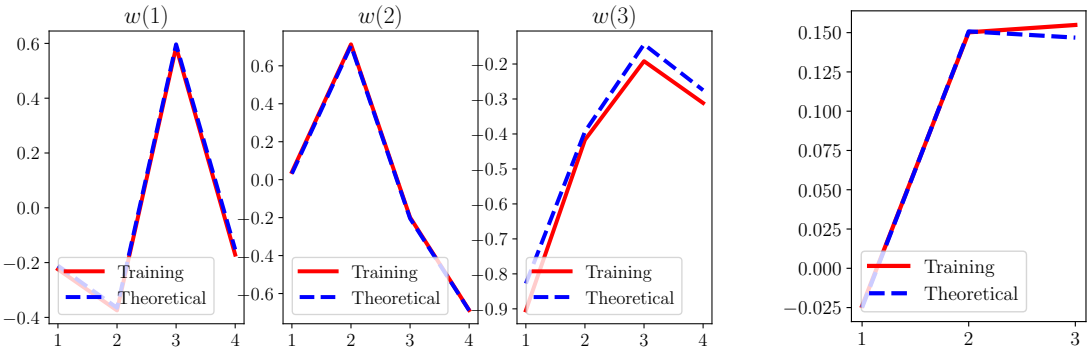
图 3.7　Softmax 回归中特征及权重的训练结果 (实线) 与理论值 (虚线) 的对比

P_{XY}，并由该分布生成 $n = 100,000$ 个 (X, Y) 的训练样本。由这些样本训练网络之后所得 $f(x), g(y)$ 及 $b(y)$ 与相应理论值的对比结果如图 3.7 所示，从中可看出两者的一致性。

基于相同的训练数据，可对定理 3.5 的结果进行检验。具体地，这里采用图 3.2 所示神经网络，并设置隐层神经元数分别为 $m = 4$、$k = 3$。使用 X 的随机函数作为输入 $t(X)$，并以 Sigmoid 函数作为激活函数 $\sigma(\cdot)$。在网络训练过程中，固定输出层的权重与偏置项，仅训练隐层的权重 $w(1), w(2), w(3)$ 及偏置项 c 以优化 Log 损失函数。相应结果如图 3.8 所示，从中可看出仿真结果与理论值的一致性。

3.7.2　神经网络性能度量

进一步地，设计实验考察神经网络分类准确率与所提取特征的 H 评分函数的关系。为此，在 ILSVRC2012 竞赛[47] 所使用的 ImageNet 图像集上进行验证，该



(a) $w(1), w(2), w(3)$　　(b) c

图 3.8　隐层中的权重与偏置项，其中实线与虚线分别表示训练结果及对应理论值

表 3.1　H 评价函数与准确率对比，其中 H_{AIC} 表示带 AIC 修正项的 H 评价函数.

模型	$H(f)$	$H_{\text{AIC}}(f)$	准确率
VGG16	148.3	41.9	0.642
VGG19	152.7	42.2	0.647
MobileNet	45.9	42.6	0.684
DenseNet121	59.5	53.3	0.714
DenseNet169	81.2	70.2	0.736
DenseNet201	89.1	73.5	0.744
Xception	179.8	162.2	0.775
InceptionV3	181.2	162.9	0.763
InceptionResNetV2	241.1	198.1	0.791

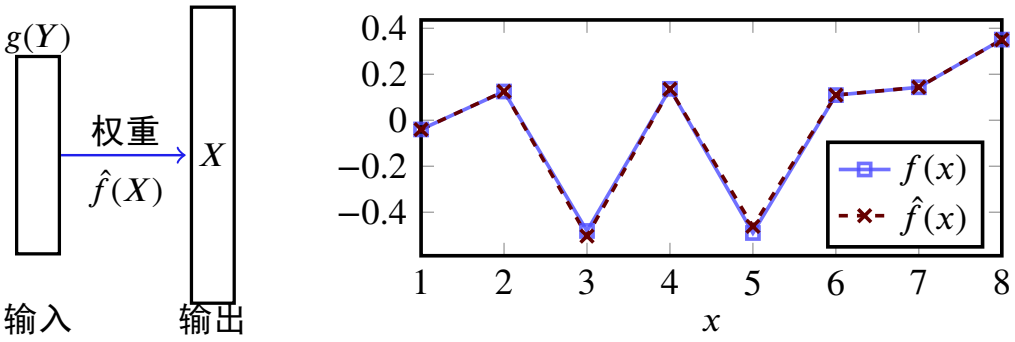
数据集包含 $n_{\text{s}} = 1,300,000$ 个训练样本。具体地，这里训练一系列常用的神经网络结构[48-53] 以获取网络最后隐层所提取图像特征，并在此基础上将分类任务准确率及网络所提取特征的 H 评分函数作对比，结果如表 3.1 所示。其中，H_{AIC} 表示根据 Akaike 信息准则 (Akaike Information Criterion, AIC)[54] 修正后的 H 评分函数，用于修正参数带来的过拟合效应，具体定义为

$$H_{\text{AIC}}(s) = H(s) - \frac{n_{\text{p}}}{n_{\text{s}}}, \tag{3-33}$$

式中 n_{p} 表示模型参数总数。由表中结果可知，修正后的 H 评分函数与准确率趋势一致，从而检验了 H 评分函数在实际学习任务中的有效性。

3.7.3　神经网络对称性

为验证命题 3.1 中神经网络对称性的结论，采用与第 3.7.1 节相同方式生成仿真数据，并以其网络结构作为这里的 X-Y 网络。训练该网络可得 (特征，权重) 二



(a) 简化的逆网络　　(b) $f(x)$ 与 $\hat{f}(x)$ 的对比

图 3.9　神经网络对称性检验

元组 (f, g)。

接着，使用 $g(y_i)$ 作为图 3.9(a) 所示逆网络[①] 的输入特征用于预测标签 x_i，并通过训练得到相应的权重 $\hat{f}$。如图 3.9(b) 所示，$\hat{f}(x)$ 与原特征 $f(x)$ 一致，由此检验了这两个网络的对称性。

3.8　本章小结

在局部信息几何分析框架下，本章对深度神经网络的特征提取问题开展分析，给出了神经网络提取特征在统计推断上的最优性。同时，本章分析结果展示了神经网络特征提取问题的奇异值分解数学结构，并由此建立了信息论角度对神经网络的性能度量。该奇异值分解结构及性能度量将作为后续章节中，对神经网络样本复杂度及训练过程中的超参数选择等问题进一步分析的理论基础。

① 这里使用简化的逆网络替代图 3.6(b) 中所示 Y-X 网络，以强制逆网络中输入为 $g(Y)$，从而避免 (3-20) 解不唯一带来的影响。

第 4 章　样本复杂度分析

4.1　本章引言

基于第 3 章的讨论可知，在 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 弱相关的情形下，深度神经网络的特征提取过程等价于求解最大相关函数，且损失函数可通过 H 评分函数等价刻画。为建立对深度神经网络的样本复杂度的理解，本章对最大相关函数的样本复杂度进行考察。具体而言，对由给定有限样本估计得到的最大相关函数，本章通过 H 评分函数度量该估计的误差，并使用对应的误差指数作为样本复杂度的描述。由于最大相关函数计算等价于求解典型相关矩阵的奇异值分解问题，本章引入矩阵微扰分析的结果以刻画由经验分布造成的误差，并由此建立误差指数的解析表达式。我们进一步将该结论推广到半监督学习的场景中，并得到半监督学习中对有标签样本与无标签样本的最优采样策略。

本章具体内容安排如下：首先，第 4.2 节给出该样本复杂度问题的数学描述并定义了描述样本复杂度的误差指数；为求解该误差指数，第 4.3 节建立了矩阵微扰分析框架以分析最大相关函数的偏差；在此基础上，第 4.4 节与第 4.5 节分别给出了有监督学习及半监督学习问题中描述样本复杂度的误差指数，并建立了半监督学习问题中的最优采样方案。最后，第 4.6 节介绍了相应的仿真实验以检验理论结果，第 4.7 节总结了全章。

4.2　问题构建

对给定样本集，在局部分析机制下深度神经网络等价于计算经验分布所对应的最大相关函数，或基于经验分布的交替条件期望算法，具体过程可描述如下。给定由联合分布[①] P_{XY} 独立同分布生成的 n 个训练样本 $(x_1, y_1), \cdots, (x_n, y_n)$，为由交替求解期望算法计算定义 2.4 所定义的最大相关函数，首先取样本均值为零的函数 $\hat{f} \in \mathbb{R}^k$ 与 $\hat{g} \in \mathbb{R}^k$，并交替执行如下过程 (参照(3-11))：

$$\begin{aligned} &\text{i)} \quad \hat{f}(x) \leftarrow \mathbb{E}\left[\boldsymbol{\Lambda}_{\hat{g}}^{-1}\hat{g}(\boldsymbol{Y}) \,\middle|\, X = x\right] \\ &\text{ii)} \quad \hat{g}(y) \leftarrow \mathbb{E}\left[\boldsymbol{\Lambda}_{\hat{f}}^{-1}\hat{f}(\boldsymbol{X}) \,\middle|\, Y = y\right] \end{aligned} \tag{4-1}$$

① 这里假设边缘分布对任意 x, y 满足 $P_X(x) > 0$ 以及 $P_Y(y) > 0$，否则可剔除相应的字母。

其中 $\hat{P}_{XY}$ 表示样本的经验分布，且期望操作分别取在样本经验条件分布 $\hat{P}_{Y|X}$ 及 $\hat{P}_{X|Y}$ 上。此外，矩阵 $\mathbf{\Lambda}_{\hat{f}}$ 及 $\mathbf{\Lambda}_{\hat{g}}$ 为协方差矩阵，定义为

$$\mathbf{\Lambda}_{\hat{f}} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_i)\hat{f}^{\mathsf{T}}(x_i), \quad \mathbf{\Lambda}_{\hat{g}} = \frac{1}{n}\sum_{i=1}^{n} \hat{g}(y_i)\hat{g}^{\mathsf{T}}(y_i).$$

在此基础上，定义诸信息向量 $\hat{\boldsymbol{\phi}}_i \in \mathbb{R}^{|\mathcal{X}|}$ 及 $\hat{\boldsymbol{\psi}}_i \in \mathbb{R}^{|\mathcal{Y}|}$，$i = 1, \cdots, k$ 使得

$$\hat{\phi}_i(x) = \sqrt{\hat{P}_X(x)}\hat{f}_i(x), \quad \hat{\psi}_i(y) = \sqrt{\hat{P}_Y(y)}\hat{g}_i(y), \tag{4-2}$$

其中经验边缘分布 $\hat{P}_X$ 与 $\hat{P}_Y$ 用作归一化因子, 且 $\hat{f}_i$ 与 $\hat{g}_i$ 分别表示 $\hat{f}$ 与 $\hat{g}$ 的第 i 维，即对任意 x, y，有

$$\hat{f}(x) = \left[\hat{f}_1(x), \cdots, \hat{f}_k(x)\right]^{\mathsf{T}}, \quad \hat{g}(y) = \left[\hat{g}_1(y), \cdots, \hat{g}_k(y)\right]^{\mathsf{T}},$$

此外，定义经验分布所对应的 $|\mathcal{Y}| \times |\mathcal{X}|$ 典型相关矩阵 $\hat{\mathbf{B}}$ 为

$$\hat{B}(y, x) = \frac{\hat{P}_{XY}(x, y)}{\sqrt{\hat{P}_X(x)\hat{P}_Y(y)}} - \sqrt{\hat{P}_X(x)\hat{P}_Y(y)}. \tag{4-3}$$

利用命题 2.3 的结论可将交替条件期望算法的执行过程 (4-1) 等价表示为

$$\begin{aligned} &\text{i)} \quad \hat{\mathbf{\Phi}}_k \leftarrow \hat{\mathbf{B}}^{\mathsf{T}}\hat{\mathbf{\Psi}}_k\left(\hat{\mathbf{\Psi}}_k^{\mathsf{T}}\hat{\mathbf{\Psi}}_k\right)^{-1} \\ &\text{ii)} \quad \hat{\mathbf{\Psi}}_k \leftarrow \hat{\mathbf{B}}\hat{\mathbf{\Phi}}_k\left(\hat{\mathbf{\Phi}}_k^{\mathsf{T}}\hat{\mathbf{\Phi}}_k\right)^{-1} \end{aligned} \tag{4-4}$$

其中

$$\hat{\mathbf{\Phi}}_k = \left[\hat{\boldsymbol{\phi}}_1, \cdots, \hat{\boldsymbol{\phi}}_k\right], \quad \hat{\mathbf{\Psi}}_k = \left[\hat{\boldsymbol{\psi}}_1, \cdots, \hat{\boldsymbol{\psi}}_k\right]. \tag{4-5}$$

注意到 (4-4) 等就与求解低秩恢复问题

$$\min_{\hat{\mathbf{\Psi}}_k, \hat{\mathbf{\Phi}}_k} \left\|\hat{\mathbf{B}} - \hat{\mathbf{\Psi}}_k\hat{\mathbf{\Phi}}_k^{\mathsf{T}}\right\|_{\mathrm{F}}^2. \tag{4-6}$$

的交替最小二乘法[55]，故由 Eckart–Young–Mirsky 定理[56] 可知 ACE 算法将收敛至 $\hat{\mathbf{B}}$ 对应于前 k 个奇异值的奇异向量。

在下面的分析中，令 $\hat{\boldsymbol{\phi}}_i$ 表示 $\hat{\mathbf{B}}$ 第 i 个右奇异向量，并令 $\hat{\mathbf{\Phi}}_k$ 表示由前 k 个右奇异向量构成的 $|\mathcal{X}| \times k$ 矩阵。注意到由命题 2.2可知最大相关函数对应于 X 和 Y 的典型相关矩阵 (参考定义 2.3) $\tilde{\mathbf{B}}$ 的前 k 个奇异向量，故当经验分布 $\hat{P}_{XY}$ 与真实分布 P_{XY} 一致时，可得 $\hat{\mathbf{B}} = \tilde{\mathbf{B}}$，从而可由交替条件期望算法得到精确的最大相关

函数值。但因为训练样本的随机性，经验分布将很大概率偏离真实分布，从而导致对应奇异向量及最大相关函数与理论值的偏差。为了量化该偏差，定义 $|\mathcal{X}| \times k$ 矩阵 $\boldsymbol{\Phi}_k$ 为

$$\boldsymbol{\Phi}_k \triangleq \left[\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_k\right],$$

其中 $\boldsymbol{\phi}_i$ 表示 $\tilde{\mathbf{B}}$ 的第 i 个右奇异向量。

在此基础上，采用 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2$ 作为 $\hat{\boldsymbol{\Phi}}_k$ 偏离 $\boldsymbol{\Phi}_k$ 程度的度量。注意该度量等价于 H 评分函数的差，从而可表示深度神经网络泛化误差与理论最优泛化误差之间的差值。

注意到对所有满足 $\hat{\boldsymbol{\Phi}}_k^{\mathsf{T}}\hat{\boldsymbol{\Phi}}_k = \mathbf{I}_k$ 的 $|\mathcal{X}| \times k$ 矩阵都有 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 \geqslant 0$，其中 $\mathbf{I}_k$ 表示 $k \times k$ 单位阵。为分析泛化误差随 n 的变化，在接下来的分析中考察误差指数[57]

$$\mathrm{E}_k \triangleq -\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_n \left\{ \left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon \right\}, \tag{4-7}$$

其中概率空间定义在所有由 P_{XY} 生成的所有可能独立同分布样本上。具体而言，(4-7) 中第一个有关 n 的极限表示 n 的大样本机制；在该机制下，经验分布与真实分布以大概率十分接近，从而 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2$ 与 $\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2$ 的差很小，故自然引入第二个关于 ϵ 的极限。

在接下来的推导中将可以看到，这两个极限的引入将大大简化样本复杂度的求解。为给出 (4-7) 的解析表达式，需刻画 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2$ 与 $\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2$ 与经验分布扰动的关系。为此，下一节将引入矩阵微扰分析的若干结果。

4.3　矩阵微扰分析

设 $\mathbf{A} \in \mathbb{R}^{d \times d}$ 为对称阵，对应特征向量与特征值分别为 $\mathbf{v}_1, \cdots, \mathbf{v}_d$ 及 $\lambda_1 \geqslant \cdots \geqslant \lambda_d$。由 $\mathbf{A}$ 的前 k 个特征向量构造矩阵

$$\mathbf{V}_k \triangleq \left[\mathbf{v}_1, \cdots, \mathbf{v}_k\right] \in \mathbb{R}^{d \times k}, \tag{4-8}$$

则可验证

$$\mathrm{tr}\left\{\mathbf{V}_k^{\mathsf{T}}\mathbf{A}\mathbf{V}_k\right\} = \sum_{i=1}^{k} \lambda_i, \tag{4-9}$$

其中 tr{·} 表示矩阵的迹。进一步地，设 $\mathbf{A}(\tau)$ 为由参数 τ 描述的一族对称矩阵，满足 $\mathbf{A}(0)=\mathbf{A}$ 且关于 τ 解析。则 $\mathbf{A}(\tau)$ 的 Taylor 级数展开可表示为

$$\mathbf{A}(\tau)=\mathbf{A}+\tau\mathbf{A}'+o(\tau),$$

其中 $\mathbf{A}'=\mathbf{A}'(0)$ 为 $\mathbf{A}(\tau)$ 关于 τ 在 $\tau=0$ 处的一阶导数。类似于(4-8)的定义，令 $\mathbf{V}_k(\tau)\in\mathbb{R}^{d\times k}$ 为由 $\mathbf{A}(\tau)$ 前 k 个特征向量构成的矩阵。那么，当 $\lambda_k>\lambda_{k+1}$ 时，以下引理给出了 $\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_k(\tau)\right\}$ 关于 τ 的二阶 Taylor 展开。

引理 4.1： 若 $\lambda_k>\lambda_{k+1}$，则

$$\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_k(\tau)\right\}=\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}\mathbf{A}\mathbf{V}_k\right\}-\tau^2\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left(\mathbf{v}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_i-\lambda_j}+o(\tau^2),$$

其中 tr{·} 表示矩阵的迹。

证明 参见附录 B.1。 □

此外，对 $\lambda_k=\lambda_{k+1}$ 的情况，我们引入记号 $[d]\triangleq\{1,\cdots,d\}$，并定义指标集 $\mathcal{I}_k\triangleq\{i\in[d]\colon\lambda_i=\lambda_k\}$ 与相应补集 $\mathcal{I}_k^{\mathsf{c}}\triangleq[d]\setminus\mathcal{I}_k=\{i\in[d]\colon\lambda_i\neq\lambda_k\}$。进一步地，定义矩阵 $\mathbf{V}_{\mathcal{I}_k}\triangleq[\mathbf{v}_i,i\in\mathcal{I}_k]\in\mathbb{R}^{d\times|\mathcal{I}_k|}$ 为 $\mathbf{V}$ 下标在 $\mathcal{I}_k$ 中的列构成的子矩阵。则对于 $\lambda_k=\lambda_{k+1}$ 的情形，如下引理给出了 $\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_k(\tau)\right\}$ 的二阶 Taylor 展开。

引理 4.2： 若 $\lambda_k=\lambda_{k+1}$，则

$$\begin{aligned}&\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_k(\tau)\right\}\\&\quad=\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}\mathbf{A}\mathbf{V}_k\right\}-\tau^2\sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\left(\mathbf{v}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_i-\lambda_j}-\tau^2\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_k-\lambda_j}+o(\tau^2),\end{aligned}$$

其中 l 为 $\mathcal{I}_k$ 中的最小元，$\hat{\mathbf{v}}_i$ 定义为

$$\hat{\mathbf{v}}_i\triangleq\mathbf{V}_{\mathcal{I}_k}\mathbf{u}_{i-l+1},\quad l\leqslant i\leqslant k,$$

其中 $\mathbf{u}_1,\cdots,\mathbf{u}_{k-l+1}\in\mathbb{R}^{|\mathcal{I}_k|}$ 为矩阵 $\mathbf{V}_{\mathcal{I}_k}^{\mathsf{T}}\mathbf{A}'\mathbf{V}_{\mathcal{I}_k}$ 的前 $k-l+1$ 个特征向量。

证明 参见附录 B.2。 □

注意到，因为 $\left\|\tilde{\mathbf{B}}\hat{\mathbf{\Phi}}_k\right\|_{\mathrm{F}}^2=\operatorname{tr}\left\{\hat{\mathbf{\Phi}}_k^{\mathsf{T}}\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\hat{\mathbf{\Phi}}_k\right\}$，本节中的结果实际上刻画了由经验分布 $\hat{P}_{XY}$ 相对于真实分布 P_{XY} 的扰动引起的 $\left\|\tilde{\mathbf{B}}\hat{\mathbf{\Phi}}_k\right\|_{\mathrm{F}}^2$ 与 $\left\|\tilde{\mathbf{B}}\mathbf{\Phi}_k\right\|_{\mathrm{F}}^2$ 之间的差异。上述矩阵微扰结果可用于后续对误差指数(4-7)的推导。

4.4　有监督学习的样本复杂度

给定由联合分布 P_{XY} 生成的 n 个独立同分布的训练样本 $(x_1, y_1), \cdots, (x_n, y_n)$，本节将分别在 $\sigma_k > \sigma_{k+1}$ 与 $\sigma_k = \sigma_{k+1}$ 两种情况下给出样本复杂度 (4-7) 的解析表达，其中 σ_k 与 σ_{k+1} 分别为 $\tilde{\mathbf{B}}$ 的第 k 与第 $(k+1)$ 大的奇异值。

4.4.1　$\sigma_k > \sigma_{k+1}$ 的情形

为了表述相应结果，我们首先引入相应的定义。首先，对给定 P_{XY}，我们定义 α_k 用于描述误差指数 (4-7)。

定义 4.1： 给定联合分布 P_{XY} 以及 $k \in \mathbb{N}^+$，定义矩阵

$$\mathbf{G}_k \triangleq \mathbf{L}^{\mathsf{T}} \left(\sum_{i=1}^{k} \sum_{j=k+1}^{d} \frac{\boldsymbol{\theta}_{ij} \boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2 - \sigma_j^2} \right) \mathbf{L}, \tag{4-10}$$

其中 $d \triangleq |\mathcal{X}|, \sigma_i$ 表示矩阵 $\tilde{\mathbf{B}}$ 的第 i 个奇异值[①]。此外，$\mathbf{L}$ 定义为 $(|\mathcal{X}| \cdot |\mathcal{Y}|) \times (|\mathcal{X}| \cdot |\mathcal{Y}|)$ 的矩阵，其第 $[(x-1)|\mathcal{Y}| + y]$ 行与第 $[(x'-1)|\mathcal{Y}| + y']$ 列的元素为

$$\sqrt{\frac{P_{XY}(x', y')}{P_X(x) P_Y(y)}} \left[\delta_{xx'} \delta_{yy'} - \frac{1}{2} \left(\frac{\delta_{xx'}}{P_X(x)} + \frac{\delta_{yy'}}{P_Y(y)} \right) \cdot \left[P_{XY}(x, y) + P_X(x) P_Y(y) \right] \right], \tag{4-11}$$

其中 δ_{ij} 为 Kronecker δ 函数，$\boldsymbol{\theta}_{ij}$ 定义为

$$\boldsymbol{\theta}_{ij} \triangleq \boldsymbol{\phi}_j \otimes \left(\tilde{\mathbf{B}} \boldsymbol{\phi}_i \right) + \boldsymbol{\phi}_i \otimes \left(\tilde{\mathbf{B}} \boldsymbol{\phi}_j \right), \quad 1 \leqslant i \leqslant j \leqslant d, \tag{4-12}$$

式中 “$\otimes$” 为矩阵 Kronecker 积，$\boldsymbol{\phi}_i$ 为 $\tilde{\mathbf{B}}$ 的第 i 个右奇异向量。接着，我们定义 $\mathbf{G}_k$ 的谱范数为 α_k。

此外，我们的分析中将会用到下列经验分布的集合。

定义 4.2： 对所有 $\epsilon > 0$，集合 $S_1(\epsilon)$ 定义为

$$S_1(\epsilon) \triangleq \left\{ \hat{P}_{XY} : \left\| \tilde{\mathbf{B}} \boldsymbol{\Phi}_k \right\|_{\mathrm{F}}^2 - \left\| \tilde{\mathbf{B}} \hat{\boldsymbol{\Phi}}_k \right\|_{\mathrm{F}}^2 > \epsilon \right\}, \tag{4-13}$$

其中 $\hat{\boldsymbol{\Phi}}_k$ 对应于 $\hat{\mathbf{B}}$ 的前 k 个右奇异向量 [参考(4-3)与(4-5)]。此外，集合 $\mathcal{N}(\epsilon)$ 定义为

$$\mathcal{N}(\epsilon) \triangleq \left\{ \hat{P}_{XY} : D(\hat{P}_{XY} \| P_{XY}) \leqslant \frac{\epsilon}{\alpha_k} \right\}. \tag{4-14}$$

① 当 $|\mathcal{X}| > |\mathcal{Y}|$ 及 $i > |\mathcal{Y}|$ 时，我们约定 $\sigma_i = 0$。

更进一步地，对于经验分布 $\hat{P}_{XY} \in \mathcal{N}(\epsilon)$，我们将 $\hat{P}_{XY}$ 与真实分布之间的 P_{XY} 差异表示为

$$\Gamma(y,x) \triangleq \begin{cases} \dfrac{\hat{P}_{XY}(x,y) - P_{XY}(x,y)}{\sqrt{\epsilon P_{XY}(x,y)}}, & \text{若 } P_{XY}(x,y) > 0, \\ 0, & \text{若 } P_{XY}(x,y) = 0, \end{cases} \tag{4-15}$$

由此建立一一对应关系[①] $\hat{P}_{XY} \leftrightarrow \Gamma$。除此之外，我们定义 $|\mathcal{Y}| \times |\mathcal{X}|$ 矩阵 $\mathbf{\Gamma}$ 及 $\mathbf{\Xi}$，使其第 y 行第 x 列的元素分别取 $\Gamma(y,x)$ 及

$$\begin{aligned} \Xi(y,x) \triangleq & \frac{\sqrt{P_{XY}(x,y)}}{\sqrt{P_X(x)P_Y(y)}}\Gamma(y,x) - \frac{P_{XY}(x,y) + P_X(x)P_Y(y)}{2\sqrt{P_X(x)P_Y(y)}} \\ & \cdot \left[\frac{1}{P_X(x)} \sum_{y' \in \mathcal{Y}} \sqrt{P_{XY}(x,y')}\Gamma(y',x) + \frac{1}{P_Y(y)} \sum_{x' \in \mathcal{X}} \sqrt{P_{XY}(x',y)}\Gamma(y,x') \right]. \end{aligned} \tag{4-16}$$

基于此，对于经验分布取自集合 $\mathcal{N}(\epsilon)$ 中的数据样本所对应的 $\hat{\mathbf{B}}$ 有如下微扰表达式。

引理 4.3： 对于给定的 P_{XY}，存在常数 $C > 0$，使得对任意 $\epsilon > 0$ 及 $\hat{P}_{XY} \in \mathcal{N}(\epsilon)$，我们有 $\|\mathbf{\Xi}\|_{\mathrm{F}} \leqslant C$ 以及

$$\hat{\mathbf{B}} = \tilde{\mathbf{B}} + \sqrt{\epsilon}\,\mathbf{\Xi} + o\left(\sqrt{\epsilon}\right). \tag{4-17}$$

证明 参见 附录 B.3. □

进一步地，如下引理刻画了 P_{XY} 往集合 $\mathcal{S}_1(\epsilon) \cap \mathcal{N}(\epsilon)$ 上的信息投影（Information Projection[58], I-projection），相应结果可用于误差指数分析。

引理 4.4： 对 $\mathcal{S}_1(\epsilon)$ 及定义 4.2中的 $\mathcal{N}(\epsilon)$，我们有[②]

$$\lim_{\epsilon \to 0^+} \epsilon^{-1} D\left(\mathcal{S}_1(\epsilon) \cap \mathcal{N}(\epsilon) \,\middle\|\, P_{XY}\right) = \frac{1}{2\alpha_k}. \tag{4-18}$$

① 注意到由于 $D(\hat{P}_{XY}\|P_{XY})$ 取有限值，对所有满足 $P_{XY}(x,y) = 0$ 的 (x,y) 我们有 $\hat{P}_{XY}(x,y) = 0$。因此，我们可以从 Γ 求得到分布 $\hat{P}_{XY}$：

$$\hat{P}_{XY}(x,y) = P_{XY}(x,y) + \sqrt{\epsilon P_{XY}(x,y)}\Gamma(y,x), \text{ 对所有 } (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

② 对于给定概率分布 P 及分布的集合 $\mathcal{S}$，我们引入记号[58]

$$D(\mathcal{S}\|P) \triangleq \inf_{Q \in \mathcal{S}} D(Q\|P).$$

证明 参见附录 B.4. □

利用引理 4.4 及 Sanov 定理[58] 的结果，可将误差指数 E_k 的解析表达式陈述为如下定理。

定理 4.1： 若 $\sigma_k > \sigma_{k+1}$，样本复杂度所对应的误差指数为

$$\mathrm{E}_k = -\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}=\frac{1}{2\alpha_k}. \tag{4-19}$$

证明 首先，根据 Sanov 定理可得

$$\mathbb{P}_n\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}\doteq\exp\left\{-nD\left(\mathcal{S}_1(\epsilon)\middle\|P_{XY}\right)\right\}. \tag{4-20}$$

因此，误差指数(4-19)可表示为

$$-\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_n\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}=\lim_{\epsilon\to 0^+}\epsilon^{-1}D\left(\mathcal{S}_1(\epsilon)\middle\|P_{XY}\right). \tag{4-21}$$

根据引理 4.4，存在 $\epsilon_0>0$ 使得对任意 $\epsilon\in(0,\epsilon_0)$,

$$D\left(\mathcal{S}_1(\epsilon)\cap\mathcal{N}(\epsilon)\,\middle\|\,P_{XY}\right)<\frac{\epsilon}{\alpha_k}.$$

此外，根据(4-14)，对所有 $\hat{P}_{XY}\in\mathcal{S}_1(\epsilon)\setminus\mathcal{N}(\epsilon)$ 我们有 $D\big(\hat{P}_{XY}\big\|P_{XY}\big)>\frac{\epsilon}{\alpha_k}$。因此，对任意 $\epsilon\in(0,\epsilon_0)$ 我们有

$$D\left(\mathcal{S}_1(\epsilon)\,\middle\|\,P_{XY}\right)=D\left(\mathcal{S}_1(\epsilon)\cap\mathcal{N}(\epsilon)\,\middle\|\,P_{XY}\right),$$

由此得到

$$\lim_{\epsilon\to 0^+}\epsilon^{-1}D\left(\mathcal{S}_1(\epsilon)\,\middle\|\,P_{XY}\right)=\lim_{\epsilon\to 0^+}\epsilon^{-1}D\left(\mathcal{S}_1(\epsilon)\cap\mathcal{N}(\epsilon)\,\middle\|\,P_{XY}\right)=\frac{1}{2\alpha_k}. \tag{4-22}$$

结合(4-21)与(4-22)的结论可推出(4-19)。 □

定理 4.1提供了学习最大相关函数的误差随样本数变化的渐进结果。而在实际应用中，更具有指导意义的是训练所需样本数的非渐进结果，如下所述。

定理 4.2： 对给定 P_{XY}，存在常数 $\epsilon_0>0$，使得对任意 $\epsilon\in(0,\epsilon_0)$ 及 $\delta\in(0,1)$，若 $n\geqslant N(\epsilon,\delta)$，则有

$$\mathbb{P}_n\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}<\delta,$$

其中

$$N(\epsilon,\delta)=\frac{4\alpha_k}{\epsilon}\left(2|\mathcal{X}||\mathcal{Y}|\log\frac{\alpha_k}{\epsilon}+\log\frac{1}{\delta}\right)=\frac{2}{\epsilon \mathrm{E}_k}\left(2|\mathcal{X}||\mathcal{Y}|\log\frac{1}{2\epsilon \mathrm{E}_k}+\log\frac{1}{\delta}\right),$$

α_k 的定义由定义 4.1给出。

证明 参见附录 B.5。 □

根据定理 4.2，为保证学习误差以不小于 $1-\delta$ 的概率小于 ϵ，只需使用 $n=O\left(\frac{1}{\epsilon \mathrm{E}_k}\log\frac{1}{\epsilon\delta \mathrm{E}_k}\right)$ 个独立样本训练最大相关函数。

当学习目标为 X,Y 的全部相关结构，即 $k=d-1$ 时，误差指数有如下的简洁表达。

推论 4.1： 若 $d=|\mathcal{X}|\leqslant|\mathcal{Y}|$，$\sigma_{d-1}>0$ 且 $k=d-1$，则可得 $\alpha_k=\frac{\sigma_1^2}{4}$ 以及

$$\mathrm{E}_k=\frac{2}{\sigma_1^2}.$$

证明 参见附录 B.6. □

4.4.2　$\sigma_k=\sigma_{k+1}$ 的情形

该情况下的样本复杂度推导与 $\sigma_k>\sigma_{k+1}$ 的情形类似。为便于结果叙述，我们首先定义

$$\mathcal{I}_k\triangleq\left\{i\in[d]:\sigma_i=\sigma_k\right\},\quad l\triangleq\min\mathcal{I}_k,\quad \text{以及}\quad \boldsymbol{\Phi}_{\mathcal{I}_k}\triangleq\left[\boldsymbol{\phi}_i:i\in\mathcal{I}_k\right]\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{I}_k|}.\tag{4-23}$$

类似于第 4.4.1 节中定义的 $\mathbf{G}_k$ 及 α_k，对 $\sigma_k=\sigma_{k+1}$ 的情形我们定义矩阵 $\mathbf{J}_k$ 及 β_k 用于表述误差指数 E_k.

定义 4.3： 对给定 $\boldsymbol{\Gamma}\in\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$，我们定义矩阵 $\mathbf{J}_k(\boldsymbol{\Gamma})\in\mathbb{R}^{(|\mathcal{X}||\mathcal{Y}|)\times(|\mathcal{X}||\mathcal{Y}|)}$ 为

$$\mathbf{J}_k(\boldsymbol{\Gamma})\triangleq\mathbf{G}_{l-1}+\mathbf{L}^{\mathsf{T}}\left(\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\boldsymbol{\vartheta}_{ij}\boldsymbol{\vartheta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}\right)\mathbf{L},\tag{4-24}$$

其中 $\mathbf{G}_{l-1}$ 及 $\mathbf{L}$ 由 (4-10)–(4-11)所定义，$\mathcal{I}_k^{\mathsf{c}}\triangleq[d]\setminus\mathcal{I}_k$；对任意 i,j，$\boldsymbol{\vartheta}_{ij}$ 定义为 $\boldsymbol{\vartheta}_{ij}\triangleq\boldsymbol{\phi}_j\otimes\left(\tilde{\mathbf{B}}\boldsymbol{\varphi}_i\right)+\boldsymbol{\varphi}_i\otimes\left(\tilde{\mathbf{B}}\boldsymbol{\phi}_j\right)$, 其中 $\boldsymbol{\varphi}_i$ 定义为

$$\boldsymbol{\varphi}_i\triangleq\boldsymbol{\Phi}_{\mathcal{I}_k}\mathbf{u}_{i-l+1},\quad l\leqslant i\leqslant k,\tag{4-25}$$

式中 $\mathbf{u}_1, \cdots, \mathbf{u}_{k-l+1} \in \mathbb{R}^{|\mathcal{I}_k|}$ 为矩阵 $\boldsymbol{\Phi}_{\mathcal{I}_k}^{\mathsf{T}} \left(\tilde{\mathbf{B}}^{\mathsf{T}} \boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}} \tilde{\mathbf{B}} \right) \boldsymbol{\Phi}_{\mathcal{I}_k}$ 的前 $k-l+1$ 个特征向量，且 $\boldsymbol{\Xi}$ 的定义由 (4-16) 给出。此外，β_k 定义为如下优化问题的最优值：

$$\underset{\boldsymbol{\Gamma}}{\text{maximize}} \quad \operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Gamma}) \, \mathbf{J}_k(\boldsymbol{\Gamma}) \operatorname{vec}(\boldsymbol{\Gamma}) \tag{4-26a}$$

$$\text{subject to} \quad \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \leqslant 1, \tag{4-26b}$$

其中 vec(·) 为矩阵的向量化操作。

以下定理给出了 σ_k 与 σ_{k+1} 相等时的误差指数。

定理 4.3： 若 $\sigma_k = \sigma_{k+1}$，样本复杂度对应的误差指数为

$$\mathrm{E}_k = -\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_n \left\{ \left\| \tilde{\mathbf{B}} \boldsymbol{\Phi}_k \right\|_{\mathrm{F}}^2 - \left\| \tilde{\mathbf{B}} \hat{\boldsymbol{\Phi}}_k \right\|_{\mathrm{F}}^2 > \epsilon \right\} = \frac{1}{2\beta_k}. \tag{4-27}$$

证明 参见附录 B.7。 □

注意到在 (4-24) 中，由于 ϑ_{ij} 取值依赖于 $\boldsymbol{\Xi}$，$\mathbf{J}_k(\boldsymbol{\Gamma})$ 取值将依赖于 $\boldsymbol{\Gamma}$。因此，与定理 4.1不同，优化问题 (B-59) 的最优值无法表达为某一给定矩阵的最大特征值，在一般情况下难以有效求解。但注意到，当固定 $\mathbf{J}_k$ 时，该优化问题将退化为 $\mathbf{J}_k$ 最大特征值的求解问题。因此，对给定 $\boldsymbol{\Gamma}$ 可以计算（或更新）$\mathbf{J}_k$，接着可求解 $\mathbf{J}_k$ 的第一个特征向量用于更新 $\boldsymbol{\Gamma}$；依此类推，可完成 $\mathbf{J}_k$ 与 $\boldsymbol{\Gamma}$ 的交替更新并得到优化问题 (B-59) 的局部最优值，具体求解步骤可总结为算法 1。特别地，为更新 $\boldsymbol{\Gamma}$ （参见算法 1 的第 13–16 行），我们将 $\operatorname{vec}(\boldsymbol{\Gamma})$ 投影至 $\mathbf{J}_k$ 最大特征值所对应的特征空间，并使用学习率 η 增强更新的鲁棒性。

一般情况下，优化问题 (B-59) 不存在闭式解。然而，对特定形式的分布，相应解具有简洁的表达。

推论 4.2： 设 $d = |\mathcal{X}| \leqslant |\mathcal{Y}|$，考虑形如

$$P_{XY}(x, y) = \begin{cases} p_1 & \text{若 } x = y \\ p_2 & \text{若 } x \neq y, \end{cases}$$

的联合分布，其中概率值 p_1 与 p_2 满足 $p_1 \neq p_2$ 及 $d[p_1 + (|\mathcal{Y}| - 1)p_2] = 1$。则对所有任意 $k \in [d-1]$，我们有 $\beta_k = \frac{\sigma_1^2}{4}$ 以及

$$\mathrm{E}_k = \frac{2}{\sigma_1^2},$$

算法 1 定理 4.3 中误差指数的计算

输入： k 及学习率 η.

计算 $\mathcal{I}_k$，l，以及 $\boldsymbol{\Phi}_{\mathcal{I}_k}$。

随机选择 $\boldsymbol{\Gamma} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$，对其归一化以使其满足 $\|\boldsymbol{\Gamma}\|_{\mathrm{F}} = 1$。

repeat

　根据 (4-16) 计算 $\boldsymbol{\Xi}$。

　$\mathbf{W} \leftarrow \boldsymbol{\Phi}_{\mathcal{I}_k}^{\mathsf{T}} \left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\Phi}_{\mathcal{I}_k}$

　for $i = 1, \cdots, k-l+1$ **do**

　　$\mathbf{u}_i \leftarrow \mathbf{W}$ 的第 i 个特征向量

　　$\boldsymbol{\varphi}_{i+l-1} \leftarrow \boldsymbol{\Phi}_{\mathcal{I}_k}\mathbf{u}_i$

　end for

　根据(4-24)计算 $\mathbf{J}_k(\boldsymbol{\Gamma})$。

　$\beta_k \leftarrow \mathrm{vec}^{\mathsf{T}}(\boldsymbol{\Gamma})\,\mathbf{J}_k(\boldsymbol{\Gamma})\,\mathrm{vec}(\boldsymbol{\Gamma})$

　计算 $\mathbf{J}_k(\boldsymbol{\Gamma})$ 最大特征值对应的特征向量 $\mathbf{q}_1, \cdots, \mathbf{q}_s$。

　$\mathbf{Q} \leftarrow [\mathbf{q}_1, \cdots, \mathbf{q}_s]$

　$\mathrm{vec}(\boldsymbol{\Gamma}) \leftarrow \mathrm{vec}(\boldsymbol{\Gamma}) + \eta\mathbf{Q}\mathbf{Q}^{\mathsf{T}}\,\mathrm{vec}(\boldsymbol{\Gamma})$

　$\mathrm{vec}(\boldsymbol{\Gamma}) \leftarrow \dfrac{\mathrm{vec}(\boldsymbol{\Gamma})}{\|\mathrm{vec}(\boldsymbol{\Gamma})\|}$

until β_k 收敛。

$\mathrm{E}_k \leftarrow (2\beta_k)^{-1}$

输出： E_k.

其中

$$\sigma_1 = \cdots = \sigma_{d-1} = \frac{|p_1 - p_2|\sqrt{d}}{\sqrt{p_1 + (d-1)p_2}} \tag{4-28}$$

为所对应 $\tilde{\mathbf{B}}$ 矩阵的非零奇异值。

证明 参见附录 B.8。 □

4.4.3　关于误差指数一般趋势的评注

在机器学习问题中，另一个值得考察的性能度量为归一化 H 评分 $\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 / \left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2$，其操作意义为学习得到的最大相关函数 $\hat{\boldsymbol{\Phi}}_k$ 相比于真实值 $\boldsymbol{\Phi}_k$ 的有效性。为研究该性能度量对应的样本复杂度问题，我们定义归一化误差指数

$\hat{\mathrm{E}}_k$ 以及相应的渐进问题

$$\hat{\mathrm{E}}_k \triangleq -\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\lim_{n\to\infty}\frac{1}{n}\log \mathbb{P}_n\left\{\frac{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2}{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2}>\epsilon\right\}, \tag{4-29a}$$

可验证 $\hat{\mathrm{E}}_k$ 满足

$$\hat{\mathrm{E}}_k=\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2\cdot \mathrm{E}_k=\left(\sum_{i=1}^{k}\sigma_i^2\right)\cdot \mathrm{E}_k, \tag{4-29b}$$

其中 $\sigma_1,\cdots,\sigma_k$ 为 $\tilde{\mathbf{B}}$ 的前 k 个奇异值。

对给定训练样本集，误差指数(4-29) 及定理 4.2的结果提供了为实现数据中相关性结构的高效提取，特征维度 k 的设计准则。具体地，由于 X 与 Y 均为离散随机变量，真实分布 P_{XY} 可由训练样本的经验分布 $\hat{P}_{XY}$ 近似，从而基于 $\hat{P}_{XY}$ 所对应的 α_k 或 β_k 可用于计算归一化误差指数 $\hat{\mathrm{E}}_k$。

在实际算法设计中，考察误差指数关于不同特征维度 k 的一半趋势会更具有指导意义。不难验证，对具有特定对称结构的联合分布，归一化的误差指数 E_k 随 k 线性变化。例如，对推论 4.2 中构造的联合分布 P_{XY}，我们有 $\hat{\mathrm{E}}_k=\left(\sum_{i=1}^{k}\sigma_i^2\right)\mathrm{E}_k=k\sigma_1^2\mathrm{E}_k=2k$。另一方面，也不难构造出误差指数不随 k 单调变化的例子，而 E_k 随 k 变化的一般趋势较为复杂。

为进一步研究 $\hat{\mathrm{E}}_k$ 的特性，我们设计了相应的仿真实验。具体地，我们从概率分布空间 $\mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ 均匀采样[①]得到联合分布 P_{XY}，并考虑误差指数 (4-29) 在多次生成的 P_{XY} 上的样本平均值。图 4.1 展示了当 $|\mathcal{X}|=12$、$|\mathcal{Y}|=10$ 时，由联合分布 P_{XY} 的 10^5 个样本计算得到的误差指数的经验平均。从图中结果可发现，当 k 较小时，误差指数随特征维度 k 线性增长，而当 k 较大时将呈现出超线性关系。虽然没有给出该结果的严格数学证明，在求解最大相关函数的实际应用中，结合图 4.1 所示趋势及定理 4.2 的结果可为维度 k 的设计提供指导。

4.5　半监督学习的样本复杂度

在半监督学习的问题中，除了 n 个有标签样本 $(x_1,y_1),\cdots,(x_n,y_n)$ 外，我们还观测到 $m=nr$ 个无标签样本 $x_{n+1},\cdots,x_{n+m}$，其中 r 表示有标签与无标签样本数的比值。此外，我们假设各无标签样本为边缘分布 P_X 的独立同分布采样，且与有标签样本独立。在该设定下，可同时利用有标签样本与无标签样本两者的信息进行最大相关函数的估计。为方便表述，记无标签样本 $x_{n+1},\cdots,x_{n+m}$ 的经验分布为

① 为生成 P_{XY}，我们独立地从 $[0,1]$ 中均匀采样作为 P_{XY} 每个元素的值，然后归一化所有采样值使其和为 1。

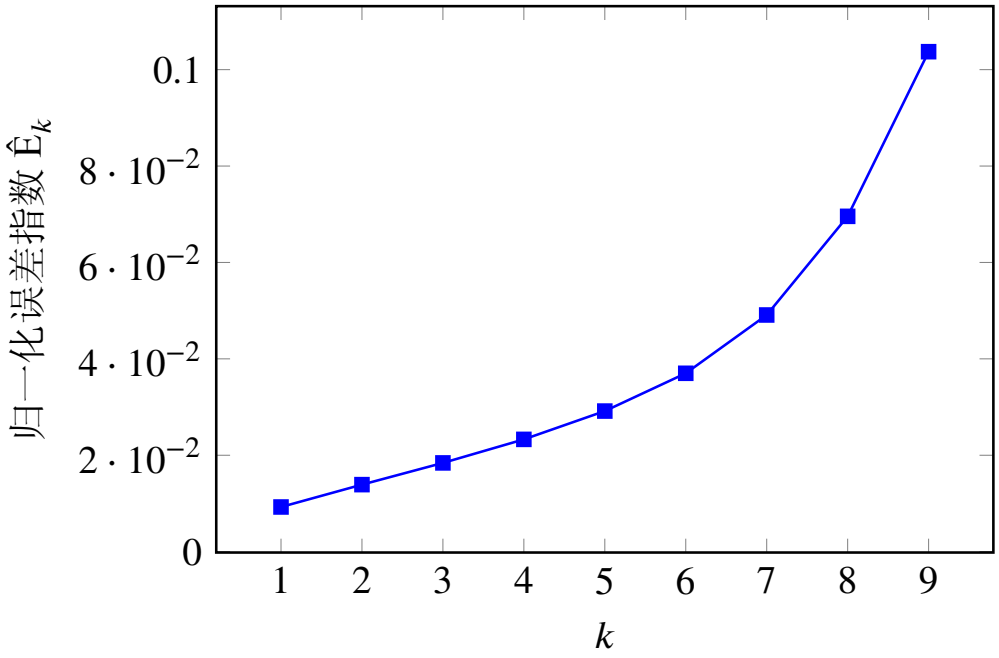


图 4.1　归一化误差指数 $\hat{\mathrm{E}}_k$ 的经验平均关于 HGR 最大相关函数维度 k 的变化趋势。该经验平均取自 10^5 个随机生成的联合分布 P_{XY}，字母集大小分别为 $|\mathcal{X}| = 12$ 及 $|\mathcal{Y}| = 10$。

Q_X，并仍采用 $\hat{P}_{XY}$ 与 $\hat{P}_{Y|X}$ 分别表示有标签数据相应的经验联合分布与经验条件分布。此外，我们将所有观测到的 X 样本 $x_1, \cdots, x_{n+m}$ 的经验分布记为 $\bar{P}_X$，则有

$$\bar{P}_X(x) = \frac{n}{m+n}\hat{P}_X(x) + \frac{m}{m+n}Q_X(x) = \frac{1}{r+1}\hat{P}_X(x) + \frac{r}{r+1}Q_X(x). \tag{4-30}$$

为了同时利用有标签与无标签样本的信息，我们可将估计最大相关函数的 ACE 算法做如下推广：

$$\begin{aligned} &\text{i)} \quad \mathbf{f}(x) \leftarrow \mathbf{\Lambda}_{\mathbf{g}}^{-1} \sum_{y\in\mathcal{Y}} \mathbf{g}(y)\hat{P}_{Y|X}(y|x) \\ &\text{ii)} \quad \mathbf{g}(y) \leftarrow \mathbf{\Lambda}_{\mathbf{f}}^{-1} \sum_{x\in\mathcal{X}} \mathbf{f}(x)\hat{P}_{Y|X}(y|x)\frac{\bar{P}_X(x)}{\bar{P}_Y(y)} \end{aligned} \tag{4-31}$$

其中：分布 $\bar{P}_Y$ 定义为

$$\bar{P}_Y(y) \triangleq \sum_{x\in\mathcal{X}} \hat{P}_{Y|X}(y|x)\bar{P}_X(x), \tag{4-32}$$

且函数 $\mathbf{f}\colon \mathcal{X} \mapsto \mathbb{R}^k$ 关于分布 $\bar{P}_X$ 均值为 0，函数 $\mathbf{g}\colon \mathcal{Y} \mapsto \mathbb{R}^k$ 关于分布 $\bar{P}_Y$ 均值为 0；$\mathbf{\Lambda}_{\mathbf{f}}$ 与 $\mathbf{\Lambda}_{\mathbf{g}}$ 为 $\mathbf{f}$ 及 $\mathbf{g}$ 的协方差矩阵，定义为

$$\mathbf{\Lambda}_{\mathbf{f}} \triangleq \sum_{x\in\mathcal{X}} \bar{P}_X(y)\mathbf{f}(x)\mathbf{f}^{\mathsf{T}}(x) = \frac{1}{n+m}\sum_{i=1}^{n+m} \mathbf{f}(x_i)\mathbf{f}^{\mathsf{T}}(x_i) \quad \text{及} \quad \mathbf{\Lambda}_{\mathbf{g}} \triangleq \sum_{y\in\mathcal{Y}} \bar{P}_Y(y)\mathbf{g}(y)\mathbf{g}^{\mathsf{T}}(y).$$

该推广算法的主要出发点是，无标签数据无法改进条件分布 $P_{Y|X}$ 的估计，但可以改进边缘分布 P_X 的估计效果。因此，算法第一步与推广前形式一致，而第二步使用了改进的经验边缘分布 $\bar{P}_X$ 以得到条件分布 $P_{X|Y}$ 更好的估计。在实践中，我们可假设从数据中估计边缘分布的难度远小于估计联合分布的难度[40]，因此该推广的 ACE 算法仍可用于样本对应最大相关函数的计算。本节目标为刻画半监督学习场景中，推广 ACE 算法 (4-31)对应的误差指数。

为此，我们定义结合有标签样本与无标签样本的联合经验分布

$$\bar{P}_{XY}(x,y) = \hat{P}_{Y|X}(y|x)\bar{P}_X(x), \tag{4-33}$$

则由附录 B.9 中推导可知算法 (4-31)实质上在计算 $|\mathcal{Y}| \times |\mathcal{X}|$ 矩阵 $\bar{\mathbf{B}}$ 的前 k 个奇异向量，其元素定义为

$$\bar{B}(y,x) = \frac{\bar{P}_{XY}(x,y)}{\sqrt{\bar{P}_X(x)\bar{P}_Y(y)}} - \sqrt{\bar{P}_X(x)\bar{P}_Y(y)}, \tag{4-34}$$

其中 $\bar{P}_Y$ 为 $\bar{P}_{XY}$ 的边缘分布，定义由 (4-32) 给出。

此外，我们记 $\bar{\mathbf{\Phi}}_k$ 为 $|\mathcal{X}| \times k$ 矩阵，使其第 i 列对应矩阵 $\bar{\mathbf{B}}$ 的第 i 个右奇异向量。于是，算法 (4-31)估计最大相关函数的样本复杂度可刻画为误差指数[59]

$$\bar{\mathrm{E}}_k(r) \triangleq -\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\lim_{n\to\infty}\frac{1}{n}\log \mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\mathbf{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\mathbf{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\}, \tag{4-35}$$

其中概率定义在由 P_{XY} 生成的 n 个独立同分布样本及由 P_X 生成的 $m = nr$ 个独立同分布样本上。类似于有监督学习中的讨论，接下来将分别在 $\sigma_k > \sigma_{k+1}$ 与 $\sigma_k = \sigma_{k+1}$ 两种情况下推导半监督学习中的误差指数(4-35)，其中 σ_k 与 σ_{k+1} 分别为矩阵 $\tilde{\mathbf{B}}$ 第 k 大与第 $(k+1)$ 大的奇异值。

4.5.1　$\sigma_k > \sigma_{k+1}$ 的情形

类似于第 4.4.1 节的讨论，我们首先定义矩阵 $\bar{\mathbf{G}}_k(r)$ 及参量 $\bar{\alpha}(r)$ 用于描指数 $\bar{\mathrm{E}}_k$ 的相关结论。

定义 4.4： 对给定 $r \geqslant 0$，定义矩阵 $\bar{\mathbf{G}}_k(r)$ 为

$$\bar{\mathbf{G}}_k(r) \triangleq \bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{G}_k\bar{\mathbf{L}}(r), \tag{4-36}$$

其中 $\mathbf{G}_k$ 定义如 (4-10) 所示，且 $\bar{\mathbf{L}}(r)$ 是维度为 $(|\mathcal{X}|\cdot|\mathcal{Y}|) \times [|\mathcal{X}|(|\mathcal{Y}|+1)]$ 的矩阵。对所有 $x, x' \in \{1,\cdots,|\mathcal{X}|\}$ 及 $y, y' \in \{1,\cdots,|\mathcal{Y}|\}$，$\bar{\mathbf{L}}(r)$ 第 $[(x-1)|\mathcal{Y}|+y]$ 行第

$[(x'-1)|\mathcal{Y}|+y']$ 列元素为

$$\delta_{xx'}\delta_{yy'}-\frac{r}{1+r}\cdot\sqrt{P_{Y|X}(y|x)P_{Y|X}(y'|x')}\cdot\delta_{xx'}, \tag{4-37a}$$

其第 $[(x-1)|\mathcal{Y}|+y]$ 行第 $[|\mathcal{X}|\cdot|\mathcal{Y}|+x']$ 列元素为

$$\frac{\sqrt{r}}{1+r}\cdot\sqrt{P_{Y|X}(y|x)}\cdot\delta_{xx'}, \tag{4-37b}$$

其中 δ_{ij} 表示 Kronecker δ 符号。在此基础上，$\bar{\alpha}_k(r)$ 定义为为矩阵 $\bar{\mathbf{G}}_k(r)$ 的谱范数。

接着我们定义联合分布 $\bar{P}_{XY}$ 相关的集合如下：

定义 4.5： 对任意 $\epsilon>0$，集合 $\bar{S}_1(\epsilon)$ 定义为

$$\bar{S}_1(\epsilon)\triangleq\left\{\bar{P}_{XY}:\ \left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}, \tag{4-38}$$

其中 $\bar{\boldsymbol{\Phi}}_k$ 对应于(4-34) 所定义的矩阵 $\bar{\mathbf{B}}$ 前 k 个右奇异向量。此外，集合 $\bar{\mathcal{N}}(\epsilon)$ 定义为

$$\bar{\mathcal{N}}(\epsilon)\triangleq\left\{\bar{P}_{XY}:\ D(\hat{P}_{XY}\|P_{XY})+rD(Q_X\|P_X)\leqslant\frac{\epsilon}{\bar{\alpha}_k(r)}\right\}, \tag{4-39}$$

其中对给定的 $\hat{P}_{XY}$ 及 Q_X，联合分布 $\bar{P}_{XY}$ 的定义由 (4-33) 给出。

其次，对任意 $\bar{P}_{XY}\in\bar{\mathcal{N}}(\epsilon)$ 及其所对应的经验分布 $\hat{P}_{XY}$ 与 Q_X，引入一一对应关系 $\hat{P}_{XY}\leftrightarrow\Gamma$ 及 $Q_X\leftrightarrow\zeta$，其中 $\Gamma(y,x)$ 定义由 (4-15)给出，ζ 类似地定义为

$$\zeta(x)\triangleq\frac{Q_X(x)-P_X(x)}{\sqrt{\epsilon P_X(x)}}. \tag{4-40}$$

再次，我们定义 $\boldsymbol{\zeta}$ 为以 $\zeta(x)$ 第 x 个元素的 $|\mathcal{X}|$ 维向量，并定义 $|\mathcal{Y}|\times|\mathcal{X}|$ 的矩阵 $\bar{\boldsymbol{\Xi}}$ 使其对应元素 $\bar{\Xi}(y,x)$ 为

$$\begin{aligned}\bar{\Xi}(y,x)\triangleq{}&\frac{\sqrt{P_{XY}(x,y)}}{\sqrt{P_X(x)P_Y(y)}}\Upsilon(y,x)-\frac{P_{XY}(x,y)+P_X(x)P_Y(y)}{2\sqrt{P_X(x)P_Y(y)}}\\&\cdot\left[\frac{1}{P_X(x)}\sum_{y'\in\mathcal{Y}}\sqrt{P_{XY}(x,y')}\Upsilon(y',x)+\frac{1}{P_Y(y)}\sum_{x'\in\mathcal{X}}\sqrt{P_{XY}(x',y)}\Upsilon(y,x')\right],\end{aligned} \tag{4-41}$$

其中我们定义

$$\Upsilon(y,x)\triangleq\Gamma(y,x)+\frac{r}{1+r}\sqrt{P_{Y|X}(y|x)}\cdot\left[\zeta(x)-\sum_{y'\in\mathcal{Y}}\sqrt{P_{Y|X}(y'|x)}\Gamma(y',x)\right]. \tag{4-42}$$

类似于引理 4.3的结论，从数据中估计得到的矩阵 $\bar{\mathbf{B}}$ 亦可表达为微扰形式，如下所述。

引理 4.5： 对给定 P_{XY} 及 $r \geqslant 0$，存在常数 $\bar{C} > 0$，使得对任意 $\epsilon > 0$ 及 $\bar{P}_{XY} \in \bar{\mathcal{N}}(\epsilon)$，我们有 $\left\|\bar{\boldsymbol{\Xi}}\right\|_{\mathrm{F}} \leqslant \bar{C}$ 以及

$$\bar{\mathbf{B}} = \tilde{\mathbf{B}} + \sqrt{\epsilon}\,\bar{\boldsymbol{\Xi}} + o\left(\sqrt{\epsilon}\right). \tag{4-43}$$

证明 参见附录 B.10。 □

此外，由 Sanov 定理可得到如下引理，其对误差指数的表达将用于后续分析中。

引理 4.6： 给定 $\epsilon > 0$，我们有

$$\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} \doteq \exp\left\{-n \cdot \inf_{\bar{P}_{XY} \in \bar{\mathcal{S}}_1(\epsilon)}\left[D(\hat{P}_{XY}\|P_{XY}) + rD(Q_X\|P_X)\right]\right\}, \tag{4-44}$$

其中 $\bar{\mathcal{S}}_1(\epsilon)$ 的定义由 (4-38)给出。

证明 参见附录 B.11. □

由 (4-44) 可知，对给定 $\epsilon > 0$，错误指数由某个加权 K-L 散度的下确界决定。当我们关注满足 $\bar{P}_{XY} \in \bar{\mathcal{N}}(\epsilon)$ 的分布 $\bar{P}_{XY}$ 时，如下引理进一步刻画了该下确界在小 ϵ 条件下的行为，相应结论将用于后续分析中。

引理 4.7： 对定义 4.5所定义的 $\bar{\mathcal{S}}_1(\epsilon)$ 及 $\bar{\mathcal{N}}(\epsilon)$，我们有

$$-\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \inf_{\bar{P}_{XY} \in \bar{\mathcal{S}}_1(\epsilon) \cap \bar{\mathcal{N}}(\epsilon)} \left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right) + rD(Q_X\|P_X)\right] = \frac{1}{2\bar{\alpha}_k(r)}. \tag{4-45}$$

证明 参见附录 B.12。 □

基于 引理 4.6 及引理 4.7的结论，可将误差指数 $\bar{\mathrm{E}}_k$ 的解析表达式总结为如下定理。

定理 4.4： 若 $\sigma_k > \sigma_{k+1}$，半监督学习样本复杂度对应的误差指数为

$$\bar{\mathrm{E}}_k(r) = -\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} = \frac{1}{2\bar{\alpha}_k(r)}. \tag{4-46}$$

证明 根据引理 4.6，误差指数 $\bar{\mathrm{E}}_k(r)$ 可写为

$$\begin{aligned}\bar{\mathrm{E}}_k(r) &= -\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}\\ &=\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)\cap\bar{\mathcal{N}}(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right].\end{aligned}\tag{4-47}$$

此外，根据引理 4.7，存在 $\epsilon_0>0$ 使得对所有 $\epsilon\in(0,\epsilon_0)$，有

$$\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)\cap\bar{\mathcal{N}}(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right]<\frac{\epsilon}{\bar{\alpha}_k(r)}.$$

进一步，注意到对所有 $\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)\backslash\bar{\mathcal{N}}(\epsilon)$，我们有 $\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right]>\frac{\epsilon}{\bar{\alpha}_k(r)}$。因此，对任意 $\epsilon\in(0,\epsilon_0)$，

$$\begin{aligned}&\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right]=\\ &\qquad\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)\cap\bar{\mathcal{N}}(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right],\end{aligned}$$

由此推出

$$\begin{aligned}&\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right]\\ &\quad=\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)\cap\bar{\mathcal{N}}(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right]=\frac{1}{2\bar{\alpha}_k(r)}.\end{aligned}\tag{4-48}$$

结合 (4-47) 及 (4-48) 可推出 (4-46)。 □

与有监督学习的情形类似，我们可得到半监督学习中最大相关函数学习的非渐进样本复杂度，如下所示。

定理 4.5： 对于给定 P_{XY} 及 $r>0$，存在常数 $\bar{\epsilon}_0>0$ 使得对任意 $\epsilon\in(0,\bar{\epsilon}_0)$ 及 $\delta\in(0,1)$，若 $n\geqslant N(\epsilon,\delta,r)$，我们有

$$\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}<\delta,$$

其中

$$\begin{aligned}\bar{N}(\epsilon,\delta,r)&=\frac{4\bar{\alpha}_k(r)}{\epsilon}\left(2|\mathcal{X}|(|\mathcal{Y}|+1)\log\frac{\bar{\alpha}_k(r)}{\epsilon}+\log\frac{1}{\delta}\right)\\ &=\frac{2}{\epsilon\bar{\mathrm{E}}_k(r)}\left(2|\mathcal{X}|(|\mathcal{Y}|+1)\log\frac{1}{2\epsilon\bar{\mathrm{E}}_k(r)}+\log\frac{1}{\delta}\right),\end{aligned}$$

$\bar{\alpha}_k(r)$ 由定义 4.4 给出。

证明 参见 附录 B.13. □

进一步地，无标签样本对学习最大相关函数的增益可由如下命题刻画。

命题 4.1： 对任意 $r \geqslant 0$，由定义 4.4 给出的 $\bar{\alpha}_k(r)$ 为 r 的单调非增凸函数且满足

$$\frac{1}{1+r}\bar{\alpha}_k(0) \leqslant \bar{\alpha}_k(r) \leqslant \frac{1}{1+r}\bar{\alpha}_k(0) + \frac{r}{1+r}\bar{\alpha}_k(\infty), \tag{4-49}$$

其中 $\bar{\alpha}_k(\infty)$ 定义为[①] $\bar{\alpha}_k(\infty) \triangleq \lim_{r\to+\infty} \bar{\alpha}_k(r)$。此外，可验证 $\bar{\alpha}_k(0) = \alpha_k$，其中 α_k 定义由定义 4.1 给出。

证明 参见附录 B.14。 □

由命题 4.1 可知半监督学习中的误差指数 $\bar{\mathrm{E}}_k(r) = [2\bar{\alpha}_k(r)]^{-1}$ 为 r 的非减函数，因此结合无监督数据样本进行最大相关函数训练将获得性能提升。此外，由 (4-49) 的第一个不等式可立即推出

$$\bar{\mathrm{E}}_k(r) \leqslant (1+r)\mathrm{E}_k, \tag{4-50}$$

其中 $(1+r)\mathrm{E}_k$ 可解释为将所有 nr 个无标签数据样本替换为有标签数据样本时，即以 $n(1+r)$ 个有标签样本学习最大相关函数，所对应的样本复杂度。因此 (4-50) 中的上界表明，在一般情况下，有标签数据对估计最大相关函数的贡献比无标签数据大。但特定情形可取得该上界，意味着无标签数据可以和有标签数据起到相同的作用，如以下推论所示（对照推论 4.1的结论）。

推论 4.3： 当 $d = |\mathcal{X}| \leqslant |\mathcal{Y}|$，$\sigma_{d-1} > 0$ 及 $k = d-1$ 时我们有 $\bar{\alpha}_k(r) = \frac{\alpha_k}{1+r} = \frac{\sigma_1^2}{4(1+r)}$，因而

$$\bar{\mathrm{E}}_k(r) = (1+r)\mathrm{E}_k = \frac{2(1+r)}{\sigma_1^2}.$$

证明 参见附录 B.15。 □

在推论 4.1 及推论 4.3 中，我们研究的是 X 与 Y 整体相关结构的学习，即 $k = d-1$。在该情况下，学习前 $d-1$ 个奇异向量 $\boldsymbol{\Phi}_k = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{d-1}]$ 等价于学习最后一个奇异向量

$$\boldsymbol{\phi}_d = \left[\sqrt{P_X(1)}, \cdots, \sqrt{P_X(d)}\right]^{\mathsf{T}}.$$

注意到 $\boldsymbol{\phi}_d$ 仅仅取决于边缘分布 P_X，故此时 X 的无标签样本与 (X,Y) 的有标签样本对学习最大相关函数贡献相同，因此可取得 (4-50) 中的上界。

① 由 $\bar{\alpha}_k(r)$ 单调非增且有下界 0 可知该极限存在。

4.5.2　$\sigma_k = \sigma_{k+1}$ 的情形

在(4-23)定义的 $\mathcal{I}_k$，l 及 $\boldsymbol{\Phi}_{\mathcal{I}_k}$ 的基础上，我们进一步定义 $\bar{\beta}(r)$ 如下。

定义 4.6： 给定 $r \geqslant 0$，$\boldsymbol{\Gamma} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ 及 $\boldsymbol{\zeta} \in \mathbb{R}^{|\mathcal{X}|}$，矩阵 $\bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta})$ 定义为

$$\bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta}) \triangleq \bar{\mathbf{G}}_{l-1}(r) + \bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{L}^{\mathsf{T}} \left(\sum_{i=l}^{k} \sum_{j \in \mathcal{I}_k^{\mathsf{c}}} \frac{\bar{\boldsymbol{\vartheta}}_{ij} \bar{\boldsymbol{\vartheta}}_{ij}^{\mathsf{T}}}{\sigma_i^2 - \sigma_j^2} \right) \mathbf{L}\bar{\mathbf{L}}(r), \tag{4-51}$$

其中 $\bar{\mathbf{G}}_{l-1}$ 及 $\bar{\mathbf{L}}(r)$ 的定义由 (4-36)–(4-37)给出，$\mathbf{L}$ 的定义由 (4-11)给出；对所有 i, j，$\bar{\boldsymbol{\vartheta}}_{ij}$ 定义为 $\bar{\boldsymbol{\vartheta}}_{ij} \triangleq \boldsymbol{\phi}_j \otimes \left(\tilde{\mathbf{B}}\bar{\boldsymbol{\varphi}}_i\right) + \bar{\boldsymbol{\varphi}}_i \otimes \left(\tilde{\mathbf{B}}\boldsymbol{\phi}_j\right)$，其中 $\bar{\boldsymbol{\varphi}}_i$ 定义为

$$\bar{\boldsymbol{\varphi}}_i \triangleq \boldsymbol{\Phi}_{\mathcal{I}_k} \bar{\mathbf{u}}_{i-l+1}, \quad l \leqslant i \leqslant k, \tag{4-52}$$

式中的 $\bar{\mathbf{u}}_1, \cdots, \bar{\mathbf{u}}_{k-l+1} \in \mathbb{R}^{|\mathcal{I}_k|}$ 对应矩阵 $\boldsymbol{\Phi}_{\mathcal{I}_k}^{\mathsf{T}} \left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}} \right) \boldsymbol{\Phi}_{\mathcal{I}_k}$ 的前 $k-l+1$ 个特征向量。在此基础上，定义 $\bar{\beta}_k(r)$ 为如下优化问题的最优值：

$$\underset{\boldsymbol{\varsigma}}{\text{maximize}} \quad \boldsymbol{\varsigma}^{\mathsf{T}} \bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta}) \boldsymbol{\varsigma} \tag{4-53a}$$

$$\text{subject to} \quad \|\boldsymbol{\varsigma}\|^2 \leqslant 1, \tag{4-53b}$$

其中 $\boldsymbol{\varsigma} \in \mathbb{R}^{|\mathcal{X}|(|\mathcal{Y}|+1)}$ 定义为

$$\boldsymbol{\varsigma} \triangleq \begin{bmatrix} \operatorname{vec}(\boldsymbol{\Gamma}) \\ \sqrt{r}\boldsymbol{\zeta} \end{bmatrix}. \tag{4-54}$$

基于上述定义，样本复杂度的误差指数可由下述定理给出。

定理 4.6： 若 $\sigma_k = \sigma_{k+1}$，半监督学习中样本复杂度的误差指数为

$$\bar{\mathrm{E}}_k(r) = -\lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{n,m} \left\{ \left\| \tilde{\mathbf{B}}\boldsymbol{\Phi}_k \right\|_{\mathrm{F}}^2 - \left\| \tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k \right\|_{\mathrm{F}}^2 > \epsilon \right\} = \frac{1}{2\bar{\beta}_k(r)}. \tag{4-55}$$

证明 参见附录 B.16。 □

注意到因 $\bar{\boldsymbol{\vartheta}}_{ij}$ 取值依赖于 $\bar{\boldsymbol{\Xi}}$，优化问题 (4-51) 中的 $\bar{\mathbf{J}}_k$ 取值将依赖于 $\boldsymbol{\varsigma}$，故与定理 4.4 不同，(4-53) 的最优值不能简单写作某个给定矩阵的最大奇异值。然而对给定的 $\bar{\mathbf{J}}_k$，该优化问题可退化为解矩阵 $\bar{\mathbf{J}}_k$ 的最大奇异值。基于此，类似于第 4.4.2 节所介绍的方法, 我们可以交替求解 $\boldsymbol{\varsigma}$ 及 $\bar{\mathbf{J}}_k$ 的最优值，并将计算过程总结为算法 2。

类似推论 4.2 的结论，对特定形式的联合分布我们可以计算得到样本复杂度的闭式解。

算法 2 定理 4.6 中误差指数的计算

输入: $k, r,$ 及学习率 η

计算 $\mathcal{I}_k, l,$ 及 $\boldsymbol{\Phi}_{\mathcal{I}_k}$

随机选择某个满足 $\|\boldsymbol{\varsigma}\|^2 = 1$ 的 $\boldsymbol{\varsigma} \in \mathbb{R}^{|\mathcal{X}|(|\mathcal{Y}|+1)}$

repeat

　根据 (4-54) 计算 $\boldsymbol{\Gamma}$ 及 $\boldsymbol{\zeta}$

　根据 (4-41) 计算 $\bar{\boldsymbol{\Xi}}$

　$\mathbf{W} \leftarrow \boldsymbol{\Phi}_{\mathcal{I}_k}^{\mathsf{T}} \left(\tilde{\mathbf{B}}^{\mathsf{T}} \bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}} \tilde{\mathbf{B}} \right) \boldsymbol{\Phi}_{\mathcal{I}_k}$

　for $i = 1, \cdots, k - l + 1$ **do**

　　$\bar{\mathbf{u}}_i \leftarrow \mathbf{W}$ 的第 i 个特征向量

　　$\bar{\boldsymbol{\varphi}}_{i+l-1} \leftarrow \boldsymbol{\Phi}_{\mathcal{I}_k} \bar{\mathbf{u}}_i$

　end for

　根据(4-51)计算 $\bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta})$

　$\bar{\beta}_k(r) \leftarrow \boldsymbol{\varsigma}^{\mathsf{T}} \bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta}) \boldsymbol{\varsigma}$

　计算矩阵 $\bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta})$ 最大特征值对应的特征向量 $\mathbf{q}_1, \cdots, \mathbf{q}_s$

　$\mathbf{Q} \leftarrow \left[\mathbf{q}_1, \cdots, \mathbf{q}_s \right]$

　$\boldsymbol{\varsigma} \leftarrow \boldsymbol{\varsigma} + \eta \mathbf{Q}\mathbf{Q}^{\mathsf{T}} \boldsymbol{\varsigma}$

　$\boldsymbol{\varsigma} \leftarrow \boldsymbol{\varsigma} / \|\boldsymbol{\varsigma}\|$

until $\bar{\beta}_k(r)$ 收敛

$\bar{\mathrm{E}}_k(r) \leftarrow \left[2\bar{\beta}_k(r) \right]^{-1}$

输出: $\bar{\mathrm{E}}_k(r)$.

推论 4.4:　对推论 4.2 所构造的联合分布 P_{XY}，其所对应的 $\tilde{\mathbf{B}}$ 矩阵的所有非零奇异值满足 $\sigma_1 = \sigma_2 = \cdots = \sigma_{d-1}$，取值由 (4-28) 给出。则对任意 $k \in [d-1]$，我们有 $\bar{\beta}_k(r) = \frac{\sigma_1^2}{4(1+r)}$，且对应误差指数为

$$\bar{\mathrm{E}}_k(r) = \frac{2(1+r)}{\sigma_1^2}.$$

证明 参见附录 B.17。 □

4.5.3　总成本约束下的最优采样策略

在半监督学习中，虽然有标签数据对学习的贡献比无标签数据大，由于数据标记引入的成本，获取有标签数据的成本往往要远大于无标签数据。因此，理解学习任务中采样成本及算法性能的折中对算法设计至关重要。在本节中，我们基于最大相关函数学习的任务背景开展对该折中关系的研究。

假设获取单个有标签及无标签样本的成本分别为 C_ℓ 及 C_u，且采样总预算为 C，则有标签样本数 n_ℓ 与无标签样本数 n_u 应满足约束 $n_\ell\mathsf{C}_\ell + n_u\mathsf{C}_u \leqslant \mathsf{C}$。不失一般性，这里我们仅考虑 $\sigma_k > \sigma_{k+1}$ 的情形。此时根据定理 4.4，利用这些样本估计 k 维最大相关函数的误差指数为 $\epsilon n_\ell/[2\bar{\alpha}_k(r)]$，其中 $r = n_u/n_\ell$。因此，总采样成本约束为 C 时可达到的最优误差指数为

$$\max_{n_\ell\mathsf{C}_\ell + n_u\mathsf{C}_u \leqslant \mathsf{C}} \frac{\epsilon n_\ell}{2\bar{\alpha}_k(r)} = \max_{r\geqslant 0} \frac{\epsilon\mathsf{C}}{2(\mathsf{C}_\ell + r\mathsf{C}_u)\bar{\alpha}_k(r)},$$

由此我们立即得到如下命题。

命题 4.2： 给定采样成本约束 C，使得估计 k 维最大相关函数样本复杂度最优的有标签样本数量 n_ℓ 与无标签样本数量 n_u 为

$$n_\ell = \frac{\mathsf{C}}{\mathsf{C}_\ell + r^*\mathsf{C}_u}, \quad n_u = \frac{r^*\mathsf{C}}{\mathsf{C}_\ell + r^*\mathsf{C}_u},$$

其中

$$r^* = \arg\min_{r\geqslant 0} (\mathsf{C}_\ell + r\mathsf{C}_u)\bar{\alpha}_k(r). \tag{4-56}$$

注意到最优比值 r^* 取值独立于 C，其含义为考虑采样成本时，无标签样本与有标签样本对学习任务重要性的相对比值。尽管优化问题 (4-56) 既无解析解也非凸优化问题，我们可以通过数值微分方法[60] 求得其局部最优值。具体地，r 的局部最优值可通过如下更新准则求解：

$$r \leftarrow r - \frac{\eta}{h}\left[(\mathsf{C}_\ell + (r+h)\mathsf{C}_u)\bar{\alpha}_k(r+h) - (\mathsf{C}_\ell + r\mathsf{C}_u)\bar{\alpha}_k(r)\right],$$

其中 $h > 0$ 表示计算数值微分的步长取值，$\eta > 0$ 为梯度下降方法中的学习率。

4.6　仿真结果

在本节中，我们使用数值仿真对前述理论结果进行验证。在仿真实验中，令 $|\mathcal{X}| = |\mathcal{Y}| = 4$ 并选取联合分布为

$$P_{XY}(x,y) = \begin{cases} \frac{1}{8} & \text{若 } x = y \\ \frac{1}{24} & \text{若 } x \neq y. \end{cases} \tag{4-57}$$

具体地，我们计算估计 $k = 2$ 维最大相关函数的经验误差指数 (4-7) 及 (4-35)，并与相应的理论结果比较。注意到 (4-57) 中的联合分布 P_{XY} 为推论 4.2 及推论 4.4 所讨论情况的一个特例, 我们可以利用推论得到作为基准的理论值。

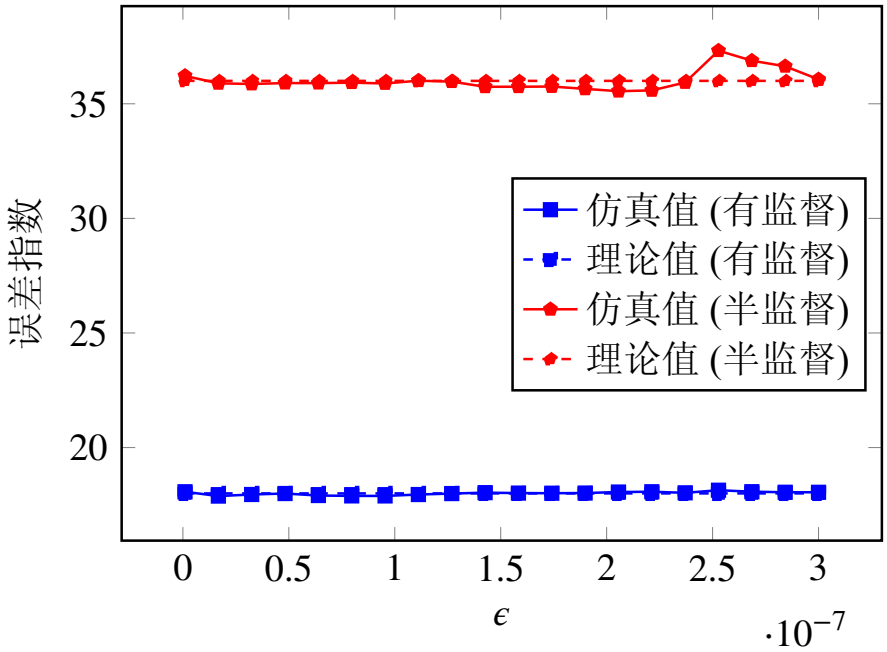


图 4.2　在有监督与半监督场景下，误差指数理论值与仿真值的比较

4.6.1　有监督学习

在本实验中，我们按如下方式采样学习误差 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2$：为采得学习误差的单个样本，我们首先由 P_{XY} 生成 $n=10^6$ 对独立同分布的 (x_i,y_i)，并根据这 n 对数据的经验分布 $\hat{P}_{XY}$ 计算 $\hat{\mathbf{B}}$。紧接着，计算 $\hat{\mathbf{B}}$ 的奇异向量以得到 $\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2$ 的一个样本。重复这样的采样过程 10^5 次，并基于这 10^5 个样本计算对应的经验错误概率

$$p_n(\epsilon)=\mathbb{P}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2>\epsilon\right\}.$$

在此基础上，可求得经验误差指数

$$-\frac{1}{n\epsilon}\log p_n(\epsilon).$$

该误差指数与由推论 4.2 计算所得理论值的对比如图 4.2 所示，从中可看出实验数据与理论结果两者相吻合。

4.6.2　半监督学习

在半监督学习的实验中，在每次采样学习误差时，我们选取 $r=1$ 并从 P_{XY} 生成 $n=10^6$ 个独立同分布的 (x_i,y_i) 对，再从 P_X 生成 $m=nr=10^6$ 个独立同分布的 x_j，接着利用 (4-33) 从数据经验分布 $\bar{P}_{XY}$ 中计算矩阵 $\bar{\mathbf{B}}$。重复这样的仿真过程

10^5 次，并计算学习误差超过 ϵ 的经验概率

$$\bar{p}_{n,m}(\epsilon) = \mathbb{P}\left\{ \left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon \right\}.$$

在此基础上，可计算得经验误差指数

$$-\frac{1}{n\epsilon}\log \bar{p}_{n,m}(\epsilon).$$

图 4.2 给出了仿真所得经验误差指数与推论 4.4 中对应的理论值的比较，从中可看出实验结果与理论的一致性。

4.7　本章小结

本章通过对最大相关函数计算误差对应的误差指数的刻画，考察了最大相关函数计算的样本复杂度。该结果也可视为是神经网络特征提取问题样本复杂度在局部分析机制下的特例。基于最大相关函数计算的与奇异值分解的本质联系，通过对矩阵奇异向量的微扰刻画，我们分别给出了有监督学习问题与半监督学习问题下的误差指数的解析表达式。基于半监督学习误差指数的结果，进一步分析了有标签样本与无标签样本对训练的贡献，并由此设计了最优的采样机制。本章的分析结果可用于指导实际学习问题中样本数设计及半监督学习中的最优采样机制设计。

第 5 章　计算效率与泛化误差最优折中

5.1　本章引言

由上一章的讨论，可得到局部分析机制下神经网络特征提取问题的泛化误差与样本数的理论关系。然而，受计算与存储资源的限制，该样本复杂度很难在实践中达到。由于实际数据往往具有高维及连续的属性，对应相当大的字母集，因此直接计算经验分布开销巨大。为在实际数据中完成该计算过程，通常做法是在每次计算中只使用一个小批量的训练集，通过随机梯度下降方法更新参数，以借助神经网络框架间接完成条件期望操作。由于样本具有随机性，随机梯度下降的超参数 (如学习率等) 的选择将直接影响算法的计算效率与最终的泛化误差。本章考察计算效率与泛化误差满足的折中关系，并进一步刻画随机梯度下降算法中的超参数选择对泛化误差的影响。

本章具体内容安排如下：首先，第 5.2 节中介绍了鲁棒条件期望算法，作为神经网络随机梯度下降方法的抽象模型，并基于该算法建立了最大相关函数计算过程中渐进泛化误差与计算效率的折中关系，以及该关系在理解残差神经网络结构中的应用。在此基础上，第 5.3 节将结论推广到一般的特征问题求解中，对应于经典的 Oja 算法的分析。具体地，第 5.3 节研究了 Oja 算法在大样本小学习率机制下的最优学习率选择，并讨论了小批量大小 (Minibatch Size) 对泛化误差的影响。接着，第 5.4 节介绍了仿真实验的结果，以检验前述理论。最后，第 5.5 节对全章内容作了小节。

5.2　鲁棒交替条件期望算法分析

首先，我们介绍求解定义 2.4 中最大相关函数的鲁棒交替条件期望算法，其可视为是对训练深度神经网络的梯度下降法在局部分析机制下的简化模型。为便于讨论，本章仅讨论 $k = 1$ 的情形，即一维最大相关函数求解。

5.2.1　鲁棒交替条件期望算法

当 $k = 1$ 时，交替条件期望算法(3-11)可表示为

$$\text{i)} \quad \boldsymbol{\phi} \leftarrow \tilde{\mathbf{B}}^{\mathsf{T}} \boldsymbol{\psi}$$

$$\text{ii)} \quad \boldsymbol{\psi} \leftarrow \tilde{\mathbf{B}} \boldsymbol{\phi}$$

其将收敛至 $|\mathcal{Y}|\times|\mathcal{X}|$ 维典型相关矩阵 $\tilde{\mathbf{B}}$ (参考定义 2.3) 的最大奇异值对应的奇异向量。若只关注 $\boldsymbol{\phi}$ 的取值，则更新规则可表示为 $\boldsymbol{\phi}\leftarrow\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}$，或等价地表示为 $f(X)\leftarrow\mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]|X\right]$。在实际应用中，由于样本数十分有限，条件期望难以精确地计算，从而直接影响计算结果。为增强结果的鲁棒性，定义鲁棒交替条件期望算法[61] 如下：

$$\boldsymbol{\phi}_n=(\mathbf{I}+\eta\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}})\boldsymbol{\phi}_{n-1},\quad n=1,2,\cdots \tag{5-1}$$

其中 $\mathbf{I}$ 为 $|\mathcal{X}|$ 阶单位阵，$\eta>0$ 为学习率，$\boldsymbol{\phi}_n$ 为第 n 次迭代所得输出。该算法可视为流式主成分分析 (Streaming Principal Component Analysis, Streaming PCA)[18-21] 中所用的 Oja 算法[32] 的一个特例。引入条件期望计算的误差后，(5-1) 可写为

$$\boldsymbol{\phi}_n=\left(\mathbf{I}+\eta\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_{n-1}+\eta\left\|\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}_{n-1}\right\|\boldsymbol{\xi}_{n-1},\quad\forall\, n, \tag{5-2}$$

其中假设误差项 $\left\|\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}_{n-1}\right\|\boldsymbol{\xi}_{n-1}$ 幅度正比于真值 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}_{n-1}$ 的幅度。同时，假设对所有 $n,\boldsymbol{\xi}_n$ 均与 $\boldsymbol{\phi}_0$ 独立，所有的 $\boldsymbol{\xi}_n$ 也相互独立且满足 $\boldsymbol{\Lambda}_{\boldsymbol{\xi}_n}=\boldsymbol{\Lambda}_{\boldsymbol{\xi}}$。由 第 3 章的分析结果，可将 (5-2) 视作深度神经网络训练中随机梯度下降的简化模型，其中噪声源于训练样本的随机性。

记 $\tilde{\mathbf{B}}$ 奇异值为 $\sigma_1\geqslant\cdots\geqslant\sigma_m=0$，其中 $m\triangleq|\mathcal{X}|$，且将对应右奇异向量分别记为 $\mathbf{v}_1,\cdots,\mathbf{v}_m$。则 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}$ 的特征值与特征向量分别为 σ_i^2 与 $\mathbf{v}_i$，$i=1,\cdots,m$。接下来，考察对给定初始向量 $\boldsymbol{\phi}_0$，由 (5-2) 计算所得 $\boldsymbol{\phi}_n$ 如何逼近目标向量 $\mathbf{v}_1$ 。为此定义性能度量 $\nu(\boldsymbol{\phi}_n)$ 及 $\bar{\nu}_n(\boldsymbol{\phi}_0)$ 如下：

$$\nu(\boldsymbol{\phi}_n)\triangleq\frac{\langle\boldsymbol{\phi}_n,\mathbf{v}_1\rangle^2}{\|\boldsymbol{\phi}_n\|^2},\quad\bar{\nu}_n(\boldsymbol{\phi}_0)\triangleq\frac{\mathbb{E}\left[\langle\boldsymbol{\phi}_n,\mathbf{v}_1\rangle^2|\boldsymbol{\phi}_0\right]}{\mathbb{E}\left[\|\boldsymbol{\phi}_n\|^2|\boldsymbol{\phi}_0\right]}, \tag{5-3}$$

其中 $\nu(\boldsymbol{\phi}_n)$ 反映了 $\boldsymbol{\phi}_n$ 与 $\mathbf{v}_1$ 间的夹角，而 $\bar{\nu}_n(\boldsymbol{\phi}_0)$ 可视为对 $\nu(\boldsymbol{\phi}_n)$ 的平均值。此外，定义 $\rho(\boldsymbol{\phi}_n)$ 及 $\bar{\rho}_n(\boldsymbol{\phi}_0)$ 为

$$\rho(\boldsymbol{\phi}_n)=\frac{\|\tilde{\mathbf{B}}\boldsymbol{\phi}_n\|^2}{\|\boldsymbol{\phi}_n\|^2},\quad\bar{\rho}_n(\boldsymbol{\phi}_0)=\frac{\mathbb{E}\left[\|\tilde{\mathbf{B}}\boldsymbol{\phi}_n\|^2\middle|\boldsymbol{\phi}_0\right]}{\mathbb{E}\left[\|\boldsymbol{\phi}_n\|^2\middle|\boldsymbol{\phi}_0\right]}. \tag{5-4}$$

注意到这里的 $\rho(\boldsymbol{\phi}_n)$ 对应于单边 H 评分函数，因此其分析可直接应用于神经网络泛化误差的分析。

在接下来的分析中，将引入记号如下：$\boldsymbol{\sigma}\triangleq(\sigma_1,\cdots,\sigma_m)^{\mathsf{T}},\boldsymbol{\Sigma}\triangleq\mathrm{diag}(\sigma_1,\cdots,\sigma_m)$，$\mathcal{V}\triangleq[\mathbf{v}_1,\cdots,\mathbf{v}_m]$，$[m]\triangleq\{1,\cdots,m\}$。此外，令 $\mathbf{1}\in\mathbb{R}^m$ 表示所有元素均为 1 的向量，并将 $\mathbb{R}^m$ 中第 i 个标准基向量记为 $\mathbf{e}_i$；令 $\circ$ 表示矩阵间 Hadamard 积，且对 $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_m)^{\mathsf{T}}$，令 $\boldsymbol{\alpha}^{\circ k}\triangleq(\alpha_1^k,\cdots,\alpha_m^k)^{\mathsf{T}}$ 表示 $\boldsymbol{\alpha}$ 的 k 次 Hadamard 幂。

此外，定义向量 $\boldsymbol{\tau}=(\tau_1,\cdots,\tau_m)^{\mathsf{T}}$ 使 $\tau_i \triangleq \mathbf{v}_i^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{\xi}}\mathbf{v}_i$，以及

$$\mathbf{G} \triangleq (\mathbf{I}+\eta\boldsymbol{\Sigma}^2)^2+\eta^2\boldsymbol{\tau}(\boldsymbol{\sigma}^{\circ 4})^{\mathsf{T}}. \tag{5-5}$$

5.2.2　最优折中关系

为简化分析过程，我们引入假设 $\sigma_1>\sigma_2$。为分析 (5-3) 及 (5-4) 中的 $\bar{v}_n(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n(\boldsymbol{\phi}_0)$，首先介绍如下引理。

引理 5.1： 给定 $m\times m$ 对角阵 $\mathbf{D}$ 及 m 维向量 $\mathbf{u},\mathbf{v}$ $(\|\mathbf{u}\|,\|\mathbf{v}\|>0)$, 设其元素分别为 D_{ii}, u_i, v_i, $i=1,\cdots,m$，则矩阵 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 特征值 $\lambda_1,\cdots,\lambda_m$ 所构成集合为可表示为 $\mathcal{S}_1\cup\mathcal{S}_2$，其中

$$\mathcal{S}_1 \triangleq \left\{\lambda:\ \sum_{i=1}^{m}\frac{u_iv_i}{\lambda-D_{ii}}=1\right\},$$

$\mathcal{S}_2 \triangleq \{D_{ii}:\, i\in[m]:\, (u_iv_i=0) \text{ or } (\exists j\neq i, D_{jj}=D_{ii})\}$，且任一元素 $\lambda\in\mathcal{S}_1\setminus\mathcal{S}_2$ 均为 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 的单特征值，对应于特征向量 $(\lambda\mathbf{I}-\mathbf{D})^{-1}\mathbf{u}$。此外，令 $\mathcal{I}_\lambda \triangleq \{i\in[m]:\, D_{ii}=\lambda\}$ 为 $\mathbf{D}$ 对角元中取值重复的元素所对应的下标集，则 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 可对角化若其满足 (i) $\mathcal{S}_1\cap\mathcal{S}_2=\varnothing$，以及 (ii) 对任意 $\lambda\in\mathcal{S}_2$，命题 $\sum_{i\in\mathcal{I}_\lambda}u_iv_i\neq 0$ 及命题 $\left(\sum_{i\in\mathcal{I}_\lambda}u_i^2\right)\cdot\left(\sum_{i\in\mathcal{I}_\lambda}v_i^2\right)=0$ 中至少一个成立。

证明 参见附录 C.1。 □

$\bar{v}_n(\boldsymbol{\phi}_0)$ 与 $\bar{\rho}_n(\boldsymbol{\phi}_0)$ 的渐进性质可由如下定理给出。

定理 5.1： 给定初始向量 $\boldsymbol{\phi}_0$，对 (5-2) 进行 n 次迭代后，由 (5-3) 及 (5-4) 定义的平均准确度 $\bar{v}_n(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n(\boldsymbol{\phi}_0)$ 可表示为

$$\bar{v}_n(\boldsymbol{\phi}_0)=\frac{\mathbf{e}_1^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbf{1}^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}},\ \bar{\rho}_n(\boldsymbol{\phi}_0)=\frac{(\boldsymbol{\sigma}^{\circ 2})^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbf{1}^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}, \tag{5-6}$$

其中 $\mathbf{G}$ 定义由 (5-5) 给出。此外，$\bar{v}_n(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n(\boldsymbol{\phi}_0)$ 分别线性收敛到

$$\bar{v}_\infty=\frac{\tau_1}{\lambda_1-(1+\eta\sigma_1^2)^2}\cdot\left(\sum_{i=1}^{m}\frac{\tau_i}{\lambda_1-(1+\eta\sigma_i^2)^2}\right)^{-1}$$

及

$$\bar{\rho}_\infty=\left(\sum_{i=1}^{m}\frac{\sigma_i^2\tau_i}{\lambda_1-(1+\eta\sigma_i^2)^2}\right)\cdot\left(\sum_{i=1}^{m}\frac{\tau_i}{\lambda_1-(1+\eta\sigma_i^2)^2}\right)^{-1},$$

且收敛率均为 $r=\lambda_2/\lambda_1$，其中 $\tilde{\boldsymbol{\phi}}_0 \triangleq \mathbf{V}^{\mathsf{T}}\boldsymbol{\phi}_0$ 且 λ_1,λ_2 分别为方程

$$\sum_{i=1}^{m}\frac{\eta^2\sigma_i^4\tau_i}{\lambda-(1+\eta\sigma_i^2)^2}=1. \tag{5-7}$$

的最大根及次大根。

证明 令 $\tilde{\boldsymbol{\phi}}_n \triangleq \mathbf{V}^{\mathsf{T}}\boldsymbol{\phi}_n, \tilde{\boldsymbol{\xi}}_n \triangleq \mathbf{V}^{\mathsf{T}}\boldsymbol{\xi}_n$，则易知 $\mathbb{E}[\tilde{\boldsymbol{\xi}}_n^{\circ 2}]=\boldsymbol{\tau}$, $\|\tilde{\mathbf{B}}\boldsymbol{\phi}_n\|^2=\|\boldsymbol{\Sigma}\tilde{\boldsymbol{\phi}}_n\|^2=(\boldsymbol{\sigma}^{\circ 2})^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_n^{\circ 2}$，$\|\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}_n\|^2=\|\boldsymbol{\Sigma}^2\tilde{\boldsymbol{\phi}}_n\|^2=(\boldsymbol{\sigma}^{\circ 4})^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_n^{\circ 2}$。于是 (5-2) 可表示为

$$\tilde{\boldsymbol{\phi}}_n=(\mathbf{I}+\eta\boldsymbol{\Sigma}^2)\tilde{\boldsymbol{\phi}}_{n-1}+\eta\|\boldsymbol{\Sigma}^2\tilde{\boldsymbol{\phi}}_{n-1}\|\tilde{\boldsymbol{\xi}}_{n-1}, \tag{5-8}$$

故有

$$\tilde{\boldsymbol{\phi}}_n^{\circ 2}=(\mathbf{I}+\eta\boldsymbol{\Sigma}^2)^2\tilde{\boldsymbol{\phi}}_{n-1}^{\circ 2}+\eta^2((\boldsymbol{\sigma}^{\circ 4})^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_i^{\circ 2})\tilde{\boldsymbol{\xi}}_{n-1}^{\circ 2}+2\eta\|\boldsymbol{\Sigma}^2\tilde{\boldsymbol{\phi}}_{n-1}\|(\mathbf{I}+\eta\boldsymbol{\Sigma}^2)(\tilde{\boldsymbol{\phi}}_{n-1}\circ\tilde{\boldsymbol{\xi}}_{n-1}).$$

取条件在 $\boldsymbol{\phi}_0$ 的条件期望，由诸 $\boldsymbol{\xi}_n$ 独立的假设可得

$$\begin{aligned}\mathbb{E}\left[\tilde{\boldsymbol{\phi}}_n^{\circ 2}\middle|\boldsymbol{\phi}_0\right]&=\left[(\mathbf{I}+\eta\boldsymbol{\Sigma}^2)^2+\eta^2\mathbb{E}\left[\tilde{\boldsymbol{\xi}}_{n-1}^{\circ 2}\right](\boldsymbol{\sigma}^{\circ 4})^{\mathsf{T}}\right]\mathbb{E}\left[\tilde{\boldsymbol{\phi}}_{n-1}^{\circ 2}\middle|\boldsymbol{\phi}_0\right]\\&=\mathbf{G}\mathbb{E}\left[\tilde{\boldsymbol{\phi}}_{n-1}^{\circ 2}\middle|\boldsymbol{\phi}_0\right]=\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2},\end{aligned}$$

故有

$$\bar{v}_n(\boldsymbol{\phi}_0)=\frac{\mathbb{E}[\langle\boldsymbol{\phi}_n,\mathbf{v}_1\rangle^2|\boldsymbol{\phi}_0]}{\mathbb{E}[\|\boldsymbol{\phi}_n\|^2|\boldsymbol{\phi}_0]}=\frac{\mathbf{e}_1^{\mathsf{T}}\mathbb{E}[\tilde{\boldsymbol{\phi}}_n^{\circ 2}|\boldsymbol{\phi}_0]}{\mathbf{1}^{\mathsf{T}}\mathbb{E}[\tilde{\boldsymbol{\phi}}_n^{\circ 2}|\boldsymbol{\phi}_0]}=\frac{\mathbf{e}_1^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbf{1}^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}},$$

$$\bar{\rho}_n(\boldsymbol{\phi}_0)=\frac{\mathbb{E}[\|\tilde{\mathbf{B}}\boldsymbol{\phi}_n\|^2|\boldsymbol{\phi}_0]}{\mathbb{E}[\|\boldsymbol{\phi}_n\|^2|\boldsymbol{\phi}_0]}=\frac{(\boldsymbol{\sigma}^{\circ 2})^{\mathsf{T}}\mathbb{E}[\tilde{\boldsymbol{\phi}}_n^{\circ 2}|\boldsymbol{\phi}_0]}{\mathbf{1}^{\mathsf{T}}\mathbb{E}[\tilde{\boldsymbol{\phi}}_n^{\circ 2}|\boldsymbol{\phi}_0]}=\frac{(\boldsymbol{\sigma}^{\circ 2})^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbf{1}^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}.$$

为推导 $\bar{v}_n(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n(\boldsymbol{\phi}_0)$ 的收敛结果，首先考察 $\mathbf{G}$ 的特征值分解。由引理 5.1 知，$\mathbf{G}$ 的所有特征值由 $\mathcal{S}_1\cup\mathcal{S}_2$ 给出，其中 $\mathcal{S}_1$ 由 (5-7) 的根构成，且 $\mathcal{S}_2=\{(1+\eta\sigma_{i+1}^2):i\in[m-1],(\sigma_i=\sigma_{i+1})\vee(\sigma_i=0)\}$，从而有 $\mathcal{S}_1\cap\mathcal{S}_2=\varnothing$。注意到若 $\sum_{i\in\mathcal{I}_\lambda}\sigma_i^4\tau_i=0$，有 $\lambda=1$ 及 $\sum_{i\in\mathcal{I}_\lambda}(\sigma_i^4)^2=0$ 成立，从而根据引理 5.1 结果知 $\mathbf{G}$ 可对角化。此外，由 $\mathcal{S}_1\cup\mathcal{S}_2\subset\mathbb{R}$ 知 $\mathbf{G}$ 所有特征值均为实数，可记为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m=1$，从而 $\mathbf{G}$ 的特征值分解可表示为 $\mathbf{G}=\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$，其中 $\boldsymbol{\Lambda}=\mathrm{diag}\{\lambda_1,\cdots,\lambda_m\}$。

令 $\hat{\lambda}_1,\hat{\lambda}_2$ 分别为集合 $\mathcal{S}_1$ 的最大元及次大元，则有

$$\hat{\lambda}_1>(1+\eta\sigma_1^2)^2>\hat{\lambda}_2>(1+\eta\sigma_2^2)^2\geqslant\max\mathcal{S}_2, \tag{5-9}$$

故 $\hat{\lambda}_1 = \lambda_1, \hat{\lambda}_2 = \lambda_2$。由于 $\lambda_1 \in \mathcal{S}_1 = \mathcal{S}_1 \setminus \mathcal{S}_2$，根据引理 5.1 可知 λ_1 为单特征值，对应特征向量

$$\mathbf{q}_1 \triangleq \mathbf{Q}\mathbf{e}_1 = (\lambda_1 \mathbf{I} - (\mathbf{I} + \eta\boldsymbol{\Sigma}^2)^2)^{-1}\boldsymbol{\tau}. \tag{5-10}$$

由此可得

$$\begin{aligned}\lim_{n\to\infty}(\lambda_1^{-1}\mathbf{G})^n &= \mathbf{Q}\cdot\lim_{n\to\infty}(\lambda_1^{-1}\boldsymbol{\Lambda})^n\cdot\mathbf{Q}^{-1}\\ &= \mathbf{Q}\mathbf{e}_1\mathbf{e}_1^{\mathsf{T}}\mathbf{Q}^{-1} = \mathbf{q}_1\mathbf{e}_1^{\mathsf{T}}\mathbf{Q}^{-1},\end{aligned}$$

故有

$$\bar{\nu}_\infty = \lim_{n\to\infty}\bar{\nu}_n(\boldsymbol{\phi}_0) = \lim_{n\to\infty}\frac{\mathbf{e}_1^{\mathsf{T}}(\lambda_1^{-1}\mathbf{G})^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbf{1}^{\mathsf{T}}(\lambda_1^{-1}\mathbf{G})^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}} = \frac{\mathbf{e}_1^{\mathsf{T}}\mathbf{q}_1}{\mathbf{1}^{\mathsf{T}}\mathbf{q}_1}.$$

同理可得

$$\bar{\rho}_\infty = \frac{\langle\boldsymbol{\sigma}^{\circ 2}, \mathbf{q}_1\rangle}{\langle\mathbf{1}, \mathbf{q}_1\rangle}.$$

在推导中，我们使用了假设 $\mathbf{e}_1^{\mathsf{T}}\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2} \neq 0$。从而 $\bar{\nu}_\infty$ 与 $\bar{\rho}_\infty$ 的结果可由 (5-10) 推出。

最后，注意到 λ_1, λ_2 分别为 $\mathbf{G}$ 最大及次大的特征值，收敛率的结论可从 (5-6) 中推出。 □

进一步，考虑 小学习率机制，即学习率 η 远小于 1 的情况，对应机器学习问题中的常见设定。该机制下，$\bar{\nu}_\infty$ 与 $\bar{\rho}_\infty$ 可进一步表示成 η 的一阶展开，如下所示。

定理 5.2： 在小学习率机制下，有

$$\bar{\nu}_\infty = 1 - \frac{\eta\sigma_1^4}{2}\sum_{i=2}^{m}\frac{\tau_i}{\sigma_1^2 - \sigma_i^2} + o(\eta), \tag{5-11}$$

$$\bar{\rho}_\infty = \sigma_1^2 - \frac{\eta\sigma_1^4}{2}\sum_{i=2}^{m}\tau_i + o(\eta), \tag{5-12}$$

$$r = 1 - 2\eta(\sigma_1^2 - \sigma_2^2) + o(\eta). \tag{5-13}$$

证明 考察 $\lambda_1(\eta)$ 在 $\eta \to 0$ 时的极限行为。对任意 $i > 1$，由于 $\lambda_1(\eta) > (1 + \eta\sigma_1^2)^2$, 有

$$0 < \frac{\eta^2}{\lambda_1(\eta) - (1+\eta\sigma_i^2)^2} < \frac{\eta^2}{(1+\eta\sigma_1^2)^2 - (1+\eta\sigma_i^2)^2} = \frac{1}{\sigma_1^2 - \sigma_i^2}\cdot\frac{\eta}{2 + \eta(\sigma_1^2 + \sigma_i^2)},$$

从而根据夹逼定理有

$$\lim_{\eta\to 0}\frac{\eta^2}{\lambda_1(\eta)-(1+\eta\sigma_i^2)^2}=0,\quad i>1.$$

故从 (5-7) 推出

$$\lim_{\eta\to 0}\frac{\eta^2\sigma_1^4\tau_1}{\lambda_1(\eta)-(1+\eta\sigma_1^2)^2}=1,$$

从而

$$\lambda_1(\eta)=(1+\eta\sigma_1^2)^2+\eta^2\sigma_1^4\tau_1+o(\eta^2). \tag{5-14}$$

由此可得

$$\frac{\lambda_1(\eta)-(1+\eta\sigma_1^2)^2}{\lambda_1(\eta)-(1+\eta\sigma_i^2)^2}=\frac{\eta\sigma_1^4\tau_1}{2(\sigma_1^2-\sigma_i^2)}+o(\eta),\quad i>1,$$

故

$$\begin{aligned}\bar{\nu}_\infty&=\frac{\tau_1}{\lambda_1(\eta)-(1+\eta\sigma_1^2)^2}\cdot\left(\sum_{i=1}^m\frac{\tau_i}{\lambda_1(\eta)-(1+\eta\sigma_i^2)^2}\right)^{-1}\\&=\left(1+\sum_{i=2}^m\frac{\tau_i}{\tau_1}\cdot\frac{\lambda_1(\eta)-(1+\eta\sigma_1^2)^2}{\lambda_1(\eta)-(1+\eta\sigma_i^2)^2}\right)^{-1}\\&=1-\frac{\eta\sigma_1^4}{2}\sum_{i=2}^m\frac{\tau_i}{\sigma_1^2-\sigma_i^2}+o(\eta).\end{aligned}$$

同法可得 $\bar{\rho}_\infty$ 表达式。

最后，为考察小学习率机制下收敛率 r 的行为，注意到由 (5-7) 及 (5-9) 有

$$1=\sum_{i=1}^m\frac{\eta^2\sigma_i^4\tau_i}{\lambda_2(\eta)-(1+\eta\sigma_i^2)^2}\leqslant\frac{\eta^2}{\lambda_2(\eta)-(1+\eta\sigma_2^2)^2}\sum_{i=2}^m\sigma_i^4\tau_i,$$

从而

$$0<\lambda_2(\eta)-(1+\eta\sigma_2^2)^2\leqslant\eta^2\sum_{i=2}^m\sigma_i^4\tau_i.$$

因此可得 $\lambda_2(\eta)=1+2\eta\sigma_2^2+o(\eta)$ 以及

$$r=\frac{\lambda_2(\eta)}{\lambda_1(\eta)}=\frac{1+2\eta\sigma_2^2+o(\eta)}{1+2\eta\sigma_1^2+o(\eta)}=1-2\eta(\sigma_1^2-\sigma_2^2)+o(\eta).\qquad\square$$

注释 5.1： 定理 5.2 实质上阐释了计算准确率、收敛率间的折中关系。注意到由(5-11)–(5-13)，当 $\eta \to 0$ 时，计算准确率 $\bar{\nu}_\infty$ 及 $\bar{\rho}_\infty$ 分别线性地趋于最优值 1 以及 σ_1^2, 而共同的收敛率 r 趋于 1。因此，该折中关系受 η 线性控制，可表示为

$$\bar{\nu}_\infty = 1 - \frac{\bar{r}\cdot\sigma_1^4}{4(\sigma_1^2-\sigma_2^2)}\sum_{i=2}^{m}\frac{\tau_i}{\sigma_1^2-\sigma_i^2} + o(\bar{r}),$$
$$\bar{\rho}_\infty = \sigma_1^2 - \frac{\bar{r}\cdot\sigma_1^4}{4(\sigma_1^2-\sigma_2^2)}\sum_{i=2}^{m}\tau_i + o(\bar{r}),$$

其中 $\bar{r} \triangleq 1 - r$。

此外，关于 $\boldsymbol{\phi}_n$ 的收敛性质，我们有如下概率层面的刻画。

定理 5.3： 给定初始向量 $\boldsymbol{\phi}_0$，存在正常数 η_0 及 c，使得对任意 $\eta < \eta_0$ 及 $n = \frac{1}{\sigma_1^2-\sigma_2^2}\cdot\frac{1}{\eta}\log\frac{1}{\eta}$，$\nu(\boldsymbol{\phi}_n)$ 与 $\rho(\boldsymbol{\phi}_n)$ 满足

$$\mathbb{P}\left\{\nu(\boldsymbol{\phi}_n) > 1 - \frac{c}{\nu(\boldsymbol{\phi}_0)}\cdot\frac{\eta}{\delta}\log\frac{1}{\eta}\right\} > 1-\delta, \tag{5-15}$$

以及

$$\mathbb{P}\left\{\rho(\boldsymbol{\phi}_n) > \sigma_1^2 - \frac{c\sigma_1^2}{\nu(\boldsymbol{\phi}_0)}\cdot\frac{\eta}{\delta}\log\frac{1}{\eta}\right\} > 1-\delta. \tag{5-16}$$

证明 参见附录 C.2。 □

注释 5.2： 为方便表述，本节推导中假设了 X 为离散随机变量，但相应的结论很容易推广到连续随机变量的情形。具体地，只需将矩阵乘法操作 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}$ 替代为线性算子 $\boldsymbol{B}: \mathcal{F} \mapsto \mathcal{F}$，其中 $\mathcal{F}$ 为 X 的 Borel 可测函数构成的集合，$B(f) = \mathbb{E}[\mathbb{E}[f(X)|Y]|X]$，并将噪声的协方差矩阵 $\boldsymbol{\Lambda}_{\boldsymbol{\xi}}$ 替代为相应的功率谱密度函数。

5.2.3　应用—残差学习理论解释

除 ACE 算法外，最大相关函数 $f^*(x)$ 也可直接由神经网络计算。具体地，设 x 为网络输入，$f(x)$ 为输出，训练网络参数以最大化 $\rho(\boldsymbol{\phi})$ [参见 (5-4)] ①，其中 $\boldsymbol{\phi} \leftrightarrow f$ 为对应的信息向量 (可参考定义 2.1)。在众多的网络结构设计中，实践证明带残差学习结构的网络 (如 ResNet[4]) 可高效地训练到所需最优特征。具体地，图 5.1 给

① 对给定训练样本 $(x_i, y_i), i = 1, \cdots, k$, $\rho(\boldsymbol{\phi})$ 表达式中的 $\|\boldsymbol{\phi}\|^2$ 及 $\|\tilde{\mathbf{B}}\boldsymbol{\phi}\|^2$ 分别可由 $\frac{1}{k}\sum_{i=1}^{k} f^2(x_i)$ 及 $\frac{1}{k}\sum_{i=1}^{k} h^2(y_i)$ 计算得到，其中对所有 $y \in \mathcal{Y}$，$h(y) = \sum_{i=1}^{k} f(x_i)\mathbb{1}\{y_i = y\} / \sum_{i=1}^{k}\mathbb{1}\{y_i = y\}$。从而可利用训练样本优化网络参数以最大化 $\rho(\boldsymbol{\phi}_n)$ 。

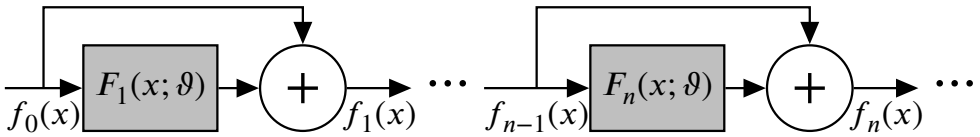


图 5.1　多层残差网络结构，其中第 n 个模块的输出为前一模块的输出及参数化函数族 $F_n(x;\vartheta)$ 的和。

出了多层残差学习的结构，其中每一模块包括相应的参数化函数族及前一模块输出的直连，每一模块通过优化参数化函数族中的参数使得模块输出可帮助后续模块学习所需特征。当模块输出特征给定时，对应参数化函数族实际学习的是该输出特征与上一层输出之间的残差。

为研究残差学习在逼近目标函数中的作用，这里考虑一种特殊的参数优化策略：每个模块均优化参数使得相应的 $\rho(\boldsymbol{\phi}_n)$ 最大化。相较于联合优化所有参数使得最终输出所对应的 $\rho(\cdot)$ 取值最大化的做法，该策略可视为分布式局部优化。如此，可建立鲁棒交替条件期望算法 (5-1) 与残差学习的联系如下。

命题 5.1： 给定输入向量 $\boldsymbol{\phi}$ 及某个较小的 $\delta>0$，在 $\|\hat{\boldsymbol{\phi}}\|\leqslant\delta$ 约束下[①]使得 $\rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})$ 最大化的 $\hat{\boldsymbol{\phi}}$ 为

$$\underset{\|\hat{\boldsymbol{\phi}}\|\leqslant\delta}{\arg\max}\ \rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})=\delta\cdot\frac{\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}-\rho(\boldsymbol{\phi})\cdot\boldsymbol{\phi}}{\|\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}-\rho(\boldsymbol{\phi})\cdot\boldsymbol{\phi}\|}+o(\delta). \tag{5-17}$$

证明 由于 $\|\hat{\boldsymbol{\phi}}\|\leqslant\delta$，可将 $\rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})$ 近似为对应的一阶 Taylor 展开

$$\begin{aligned}\rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})&=\rho(\boldsymbol{\phi})+\langle\nabla\rho(\boldsymbol{\phi}),\hat{\boldsymbol{\phi}}\rangle+o(\delta)\\&=\rho(\boldsymbol{\phi})+\frac{2}{\|\boldsymbol{\phi}\|^2}\langle\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}-\rho(\boldsymbol{\phi})\cdot\boldsymbol{\phi},\hat{\boldsymbol{\phi}}\rangle+o(\delta),\end{aligned}$$

故最大化 $\rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})$ 的 $\hat{\boldsymbol{\phi}}$ 平行于 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}-\rho(\boldsymbol{\phi})\cdot\boldsymbol{\phi}$，从而得 (5-17)。 □

注意到由 (5-17)，最优输出 $\boldsymbol{\phi}+\hat{\boldsymbol{\phi}}$ 在进行合适的线性缩放并忽略 $o(\delta)$ 项之后可表示为 $(\mathbf{I}+\eta\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}})\boldsymbol{\phi}$，其中 η 为标量。该结果建立了残差学习与迭代过程 (5-1) 的联系。若进一步假设 Y 为二元随机变量，则最大相关函数 $f^*(X)$ 正比于 $(P_{X|Y=0}(x)-P_X(x))/P_X(x)$，从而残差模块学习的目标为二元假设检验问题中基于 X 推断 Y 的似然函数。在该条件下，可不借助 δ 很小的约束直接建立残差学习与迭代算法 (5-1) 的联系，如下所述。

① δ 较小的约束实质上限制每个参数化函数族只能学习到最优特征的一小部分信息，即残差网络中每个模块只需逼近目标特征的某一部分，并使得最终输出接近目标特征。

命题 5.2： 设 $|\mathcal{Y}| = 2$，对给定输入 $\boldsymbol{\phi}$ 及 $\delta > 0$，在约束 $\|\hat{\boldsymbol{\phi}}\| \leqslant \delta$ 下最大化 $\rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})$ 的 $\hat{\boldsymbol{\phi}}$ 为

$$\underset{\|\hat{\boldsymbol{\phi}}\|\leqslant\delta}{\arg\max}\ \rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}}) = \delta \cdot \frac{\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi} - \alpha(\delta)\cdot\boldsymbol{\phi}}{\|\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi} - \alpha(\delta)\cdot\boldsymbol{\phi}\|}, \tag{5-18}$$

其中 $\alpha(\delta)$ 为某一依赖于 δ 的标量。

证明 当 $|\mathcal{Y}| = 2$ 时，矩阵 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}$ 秩为 1，因此对任意 $\boldsymbol{\phi}$，$\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}$ 均平行于 $\mathbf{v}_1$，其中 $\mathbf{v}_1$ 表示 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}$ 的最大特征值对应的特征向量。由 Lagrange 乘子法可知最优 $\hat{\boldsymbol{\phi}}$ 满足

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial\hat{\boldsymbol{\phi}}}\left(\rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}}) - \alpha\left(\|\hat{\boldsymbol{\phi}}\|^2 - \delta^2\right)\right) \\
&= \frac{2}{\|\boldsymbol{\phi}+\hat{\boldsymbol{\phi}}\|^2}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}}) - \rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})\cdot(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})\right) - 2\alpha\hat{\boldsymbol{\phi}} \\
&= \frac{2}{\|\boldsymbol{\phi}+\hat{\boldsymbol{\phi}}\|^2}\left(a\cdot\mathbf{v}_1 - \rho(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})\cdot(\boldsymbol{\phi}+\hat{\boldsymbol{\phi}})\right) - 2\alpha\hat{\boldsymbol{\phi}},
\end{aligned} \tag{5-19}$$

其中 a 与 α 均为标量。因 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi}$ 平行于 $\mathbf{v}_1$，由 (5-19) 可知最优的 $\hat{\boldsymbol{\phi}}$ 平行于 $\tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}}\boldsymbol{\phi} - \alpha(\delta)\cdot\boldsymbol{\phi}$，其中 $\alpha(\delta)$ 为与 δ 相关的标量。 □

根据命题 5.1 与命题 5.2，可利用第 5.2.2 节的结论对残差学习及相应的多模块结构的作用进行初步的理论解释。为此，注意到在图 5.1 所示的每一个残差模块中，参数化函数族将学习到最大相关函数的一小部分，对应于 (5-2) 中较小的学习率 η，相应的逼近误差对应于 (5-2) 中的噪声项。由定理 5.2 可知小学习率将有助于提高逼近最大相关函数的准确度，但相应的收敛速度也将减慢。因此，为达到这样的准确度，算法 (5-2) 将需要较多的迭代次数，对应于残差学习中的多个残差模块。为达到较好的学习效果，实践中的残差学习 (例如 ResNet) 大多基于深层神经网络。

5.3　Oja 算法分析

本节对上一节针对最大相关函数计算的结论推广至一般特征问题，并进一步考察相应的求解特征向量的算法——Oja 算法的相应性质[62]。

5.3.1　问题构建

具体地，设 $\mathbf{A} \in \mathbb{R}^{d\times d}$ 为半正定矩阵，且特征值分解为 $\mathbf{A} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}$，其中 d 为 $\mathbf{A}$ 的维度，且 $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_d]$ 及 $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_d)$ $(\sigma_1 \geqslant \cdots \geqslant \sigma_d)$ 分别为 $\mathbf{A}$ 的特征向量与特征值。于是，用于计算 $\mathbf{A}$ 第一个特征向量的 Oja 算法[32] 可表示为算

算法 3 含噪声项 $\boldsymbol{\zeta}_n$ 的 Oja 算法[32]

输入: $\boldsymbol{\phi}_0$.
for $n = 1$ **to** N **do**
　$\boldsymbol{\psi}_n \leftarrow \mathbf{A}\boldsymbol{\phi}_{n-1} + \boldsymbol{\zeta}_n$
　$\boldsymbol{\phi}_n \leftarrow \boldsymbol{\phi}_{n-1} + \eta\boldsymbol{\psi}_n$
end for
输出: $\boldsymbol{\phi}_N/\|\boldsymbol{\phi}_N\|$.

法 3, 其中 $\eta > 0$ 为学习率，且 $\boldsymbol{\zeta}_n$ 表示第 n 轮中 $\mathbf{A}\boldsymbol{\phi}_{n-1}$ 的估计噪声。进一步，这里假设 η 不随 n 变化，且每轮计算均基于新的样本，因此总迭代次数 N 等于总样本数。此外，假设噪声 $\boldsymbol{\zeta}_n$ 可表示为

$$\boldsymbol{\zeta}_n = \mathbf{Z}_n\boldsymbol{\phi}_{n-1} \tag{5-20}$$

的形式，其中诸 $\mathbf{Z}_n$ 为 $d \times d$ 矩阵，用于表示数据样本中的独立同分布噪声。具体地，设所有 $\mathbf{Z}_n$ 均与随机矩阵 $\mathbf{Z}$ 同分布，且 $\mathbf{Z}$ 所有元素均值均为零、方差为有限值。作为该模型的一个重要实例，下面对流式主成分分析进行简单介绍。

例 5.1 (流式主成分分析): 给定零均值随机向量 $\mathbf{x} \in \mathbb{R}^d$ 独立同分布的样本流 $\mathbf{x}_1, \cdots, \mathbf{x}_N$，则 $\mathbf{x}$ 的主成分可由如下更新规则计算：

$$\boldsymbol{\phi}_n \leftarrow \boldsymbol{\phi}_{n-1} + \eta\langle\boldsymbol{\phi}_{n-1}, \mathbf{x}_n\rangle\mathbf{x}_n,$$

其对应于算法 3 中 $\mathbf{A} = \mathrm{cov}(\mathbf{x})$、$\mathbf{Z}_n = \mathbf{x}_n\mathbf{x}_n^{\mathsf{T}} - \mathbf{A}$ 的情况。

首先将算法 3 的计算结果等价表示为①

$$\boldsymbol{\phi}_n = (\mathbf{I} + \eta\mathbf{A})\boldsymbol{\phi}_{n-1} + \eta\boldsymbol{\zeta}_n, \quad n = 1, \cdots, N, \tag{5-21}$$

以便于后续理论推导。与第 5.2 节中类似，这里考察给定 $\boldsymbol{\phi}_0$，$\boldsymbol{\phi}_n$ 逼近最优值 $\mathbf{v}_1$ 的行为。同样地，这里定义性能度量

$$\nu(\boldsymbol{\phi}_n) \triangleq \frac{\langle\boldsymbol{\phi}_n, \mathbf{v}_1\rangle^2}{\|\boldsymbol{\phi}_n\|^2}, \quad \rho(\boldsymbol{\phi}_n) \triangleq \frac{\boldsymbol{\phi}_n^{\mathsf{T}}\mathbf{A}\boldsymbol{\phi}_n}{\|\boldsymbol{\phi}_n\|^2}, \tag{5-22}$$

其中 $\rho(\boldsymbol{\phi}_n)$ 为 Rayleigh 商[63]。注意到 $\nu(\boldsymbol{\phi}_n)$ 及 $\rho(\boldsymbol{\phi}_n)$ 可视为对泛化性能的度量，从而相应的泛化误差可使用

$$\nu^{\mathsf{c}}(\boldsymbol{\phi}_n) \triangleq \nu(\mathbf{v}_1) - \nu(\boldsymbol{\phi}_n) = 1 - \nu(\boldsymbol{\phi}_n) = \sin^2(\boldsymbol{\phi}_n, \mathbf{v}_1)$$

① 因主成分计算由算法 3 而非该等价表达式完成，故计算成本保持不变。

$$\rho^{\mathsf{C}}(\boldsymbol{\phi}_n) \triangleq \rho(\mathbf{v}_1) - \rho(\boldsymbol{\phi}_n) = \sigma_1 - \rho(\boldsymbol{\phi}_n)$$

进行刻画。此外，定义平均意义下的性能度量

$$\bar{\nu}_n(\boldsymbol{\phi}_0) \triangleq \frac{\mathbb{E}[\langle \boldsymbol{\phi}_n, \mathbf{v}_1\rangle^2 | \boldsymbol{\phi}_0]}{\mathbb{E}[\|\boldsymbol{\phi}_n\|^2 | \boldsymbol{\phi}_0]},\ \bar{\rho}_n(\boldsymbol{\phi}_0) \triangleq \frac{\mathbb{E}[\boldsymbol{\phi}_n^{\mathsf{T}}\mathbf{A}\boldsymbol{\phi}_n | \boldsymbol{\phi}_0]}{\mathbb{E}[\|\boldsymbol{\phi}_n\|^2 | \boldsymbol{\phi}_0]}, \tag{5-23}$$

或等价地描述为平均泛化误差与其最优值的间隙

$$\bar{\nu}_n^{\mathsf{C}}(\boldsymbol{\phi}_0) \triangleq \bar{\nu}_n(\mathbf{v}_1) - \bar{\nu}_n(\boldsymbol{\phi}_0) = 1 - \bar{\nu}_n(\boldsymbol{\phi}_0) \tag{5-24a}$$

$$\bar{\rho}_n^{\mathsf{C}}(\boldsymbol{\phi}_0) \triangleq \bar{\rho}_n(\mathbf{v}_1) - \bar{\rho}_n(\boldsymbol{\phi}_0) = \sigma_1 - \bar{\rho}_n(\boldsymbol{\phi}_0). \tag{5-24b}$$

为便于表述结果，进一步引入定义如下：$\tilde{d} \triangleq d(d+1)/2$；令 $\mathbf{I}_{\tilde{d}}$ 表示 $\tilde{d}$ 阶单位阵，并记其 $[i+(j-1)j/2]$ 列为 $\mathbf{e}_{ij}$；定义 $\tilde{d}\times\tilde{d}$ 维矩阵 $\mathbf{G}$ 为

$$\mathbf{G} \triangleq \mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Sigma}_1 + \eta^2\big(\boldsymbol{\Sigma}_2 + \mathbf{L}\big), \tag{5-25}$$

其中：$\boldsymbol{\Sigma}_1$ 与 $\boldsymbol{\Sigma}_2$ 均为 $\tilde{d}\times\tilde{d}$ 阶对角阵，且其第 $[i+(j-1)j/2]$ 个对角元分别为 $\sigma_i+\sigma_j$ 及 $\sigma_i\sigma_j$；$\mathbf{L}$ 为 $\tilde{d}\times\tilde{d}$ 阶矩阵且其第 $[i+(j-1)j/2]$ 行、第 $\left[i'+(j'-1)j'/2\right]$ 列元素为

$$L_{ij,i'j'} \triangleq \mathbb{E}\left[\mathrm{tr}\left\{\mathbf{V}_{ij}\mathbf{Z}\mathbf{V}_{i'j'}\mathbf{Z}^{\mathsf{T}}\right\}\right], \tag{5-26}$$

其中对任意 $i \leqslant j \leqslant d$，有

$$\mathbf{V}_{ij} \triangleq \begin{cases} \mathbf{v}_i\mathbf{v}_i^{\mathsf{T}}, & \text{if } i = j \\ \frac{1}{\sqrt{2}}(\mathbf{v}_i\mathbf{v}_j^{\mathsf{T}} + \mathbf{v}_j\mathbf{v}_i^{\mathsf{T}}), & \text{if } i < j \end{cases}. \tag{5-27}$$

此外，记

$$\tau_i \triangleq L_{ii,11} = \mathbb{E}\big[(\mathbf{v}_i^{\mathsf{T}}\mathbf{Z}\mathbf{v}_1)^2\big] \quad i = 1, \cdots, d. \tag{5-28}$$

5.3.2　泛化误差与最优学习率

在接下来的分析中，假设对任意 $i < j$ 以及 $i' < j'$，有 (i) $\sigma_i > \sigma_j$, 以及 (ii) $\sigma_i + \sigma_j = \sigma_{i'} + \sigma_{j'}$ 当且仅当 $i = i'$, $j = j'$。另外，设 Oja 算法的输入 $\boldsymbol{\phi}_0$ 满足 $\langle \mathbf{v}_1, \boldsymbol{\phi}_0\rangle^2 > 0$，以避免退化到平凡的情形。

我们首先对 (5-23) 所定义的期望的泛化能力 $\bar{\nu}_n(\boldsymbol{\phi}_0)$ 即 $\bar{\rho}_n(\boldsymbol{\phi}_0)$ 进行刻画。根据

$$\|\boldsymbol{\phi}_n\|^2 = \sum_{i=1}^{d}\langle \boldsymbol{\phi}_n, \mathbf{v}_i\rangle^2 \quad \text{以及} \quad \boldsymbol{\phi}_n^{\mathsf{T}}\mathbf{A}\boldsymbol{\phi}_n = \sum_{i=1}^{d}\sigma_i\langle \boldsymbol{\phi}_n, \mathbf{v}_i\rangle^2,$$

可将 (5-23) 表达为

$$\bar{v}_n(\boldsymbol{\phi}_0)=\frac{\pi_n^{(1)}(\boldsymbol{\phi}_0)}{\sum_{i=1}^d \pi_n^{(i)}(\boldsymbol{\phi}_0)},\ \bar{\rho}_n(\boldsymbol{\phi}_0)=\frac{\sum_{i=1}^d \sigma_i\pi_n^{(i)}(\boldsymbol{\phi}_0)}{\sum_{i=1}^d \pi_n^{(i)}(\boldsymbol{\phi}_0)}, \tag{5-29}$$

其中 $\pi_n^{(i)}(\boldsymbol{\phi}_0)$ 定义为

$$\pi_n^{(i)}(\boldsymbol{\phi}_0)\triangleq\mathbb{E}\left[\langle\boldsymbol{\phi}_n,\mathbf{v}_i\rangle^2\middle|\boldsymbol{\phi}_0\right]. \tag{5-30}$$

进一步地，$\pi_n^{(i)}(\boldsymbol{\phi}_0)$ 满足如下命题，其证明可参见附录 C.3.

命题 5.3： 给定初始向量 $\boldsymbol{\phi}_0$，在经 (5-21) 的 n 轮迭代之后, 有

$$\pi_n^{(i)}(\boldsymbol{\phi}_0)=\langle\mathbf{e}_{ii},\mathbf{G}^n\theta_0\rangle, \tag{5-31}$$

其中 $\mathbf{e}_{ii}$ 为 $\mathbf{I}_{\tilde{d}}$ 的第 $[i(i+1)/2]$ 列, $\mathbf{G}$ 定义由 (5-25) 给出，且 $\boldsymbol{\theta}_0$ 为一 $\tilde{d}$ 维向量，其第 $[i+(j-1)j/2]$ 个元素定义为 $\boldsymbol{\phi}_0^{\mathsf{T}}\mathbf{V}_{ij}\boldsymbol{\phi}_0$ $(i\leqslant j\leqslant d)$。

注意到由命题 5.3 可立即得到关于一般的 n 的泛化误差 $\bar{v}_n^{\mathsf{G}}(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n^{\mathsf{G}}(\boldsymbol{\phi}_0)$ (或等价地，泛化性能 $\bar{v}_n(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n(\boldsymbol{\phi}_0)$) 的解析表达式。在接下来的分析中，我们将研究泛化误差 $\bar{v}_N^{\mathsf{G}}(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_N^{\mathsf{G}}(\boldsymbol{\phi}_0)$ 的理论性质，其描述了经过 N 次迭代后计算得到的 $\boldsymbol{\phi}_N$ 的准确程度。特别地，我们将考虑学习率 η 较小 (即远小于 1) 且样本数 N 较大使得乘积 $N\eta$ 较大的时算法的渐进性能，并将相应设定称为 *大样本小学习率机制*。具体地，大样本假设源于实践中的海量数据，且经验研究表明使用小的学习率将有助于降低泛化误差并且避免训练中出现参数振荡的现象[6,24-25]。此外，我们很快将表明，该机制可以给出较低泛化误差的理论保障，因此对实际应用更具有指导意义。

首先，考虑大样本小学习率机制的一个特例，并假设样本数为无穷大，即 $N=\infty$ 的情形。该特例实际上刻画了泛化误差的收敛行为。特别地，我们的结果将基于如下关于 $\pi_n^{(i)}(\boldsymbol{\phi}_0)$ 的结论，其证明参见附录 C.4。

引理 5.2： 给定小学习率 $\eta>0$，$\mathbf{G}$ 特征值为

$$\lambda_{ij}(\mathbf{G})=1+\eta(\sigma_i+\sigma_j)+\eta^2(\sigma_i\sigma_j+L_{ij,ij})+o(\eta^2),\quad 1\leqslant i\leqslant j\leqslant d, \tag{5-32}$$

且对 $i=1,\cdots,d$，有

$$\pi_n^{(i)}(\boldsymbol{\phi}_0)=(\lambda_{11}(\mathbf{G}))^n\cdot\left[\gamma_{ii}^n(\eta)\langle\mathbf{e}_{ii},\theta_0\rangle+\eta\sum_{\substack{i'\leqslant j'\\(i',j')\neq(i,i)}}\frac{\gamma_{i'j'}^n(\eta)-\gamma_{ii}^n(\eta)}{\sigma_{i'}+\sigma_{j'}-2\sigma_i}\cdot L_{ii,i'j'}\langle\mathbf{e}_{i'j'},\theta_0\rangle+o(\eta)\right] \tag{5-33}$$

其中 $L_{ii,i'j'}$ 定义由 (5-26) 给出，且 $\gamma_{ij}(\eta)$ 定义为

$$\gamma_{ij}(\eta) \triangleq \frac{\lambda_{ij}(\mathbf{G})}{\lambda_{11}(\mathbf{G})}, \qquad 1 \leqslant i \leqslant j \leqslant d. \tag{5-34}$$

基于引理 5.2，可将泛化误差 $\bar{v}_n^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 及 $\bar{\rho}_n^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 的渐进行为总结为如下定理。

定理 5.4： 给定小学习率 $\eta > 0$，, 泛化误差 $\bar{v}_n^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 与 $\bar{\rho}_n^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 分别线性收敛到

$$\bar{v}_\infty^{\mathsf{C}} = \frac{\eta}{2} \sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} + o(\eta) \tag{5-35a}$$

及

$$\bar{\rho}_\infty^{\mathsf{C}} = \frac{\eta}{2} \sum_{i=2}^{d} \tau_i + o(\eta), \tag{5-35b}$$

其共同的收敛率为 $r = 1 - \eta(\sigma_1 - \sigma_2) + o(\eta)$.

证明 对充分小的 η，由 (5-32) 可得对任意 $(i,j) \neq (1,1)$ 有 $\lambda_{ij}(\mathbf{G}) < \lambda_{11}(\mathbf{G})$，从而推出 $\gamma_{ij}(\eta) < 1$ 以及 $\lim_{n\to\infty} \gamma_{ij}^n(\eta) = 0$。故对 (5-33) 取极限，有

$$\lim_{n\to\infty} \frac{\pi_n^{(1)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^n} = \langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2 + \eta \sum_{\substack{i' \leqslant j' \\ (i',j') \neq (1,1)}} \frac{L_{11,i'j'}}{2\sigma_1 - \sigma_{i'} - \sigma_{j'}} \cdot \langle \mathbf{e}_{i'j'}, \boldsymbol{\theta}_0 \rangle + o(\eta),$$

及

$$\begin{aligned} \lim_{n\to\infty} \frac{\pi_n^{(i)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^n} &= \eta \cdot \frac{L_{ii,11} \langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2}{2(\sigma_1 - \sigma_i)} + o(\eta) \\ &= \eta \cdot \frac{\tau_i \langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2}{2(\sigma_1 - \sigma_i)} + o(\eta), \quad i > 1. \end{aligned}$$

因此，对 $i > 1$ 有

$$\begin{aligned} \lim_{n\to\infty} \frac{\pi_n^{(i)}(\boldsymbol{\phi}_0)}{\pi_n^{(1)}(\boldsymbol{\phi}_0)} &= \frac{\lim\limits_{n\to\infty} \pi_n^{(i)}(\boldsymbol{\phi}_0) \cdot (\lambda_{11}(\mathbf{G}))^{-n}}{\lim\limits_{n\to\infty} \pi_n^{(1)}(\boldsymbol{\phi}_0) \cdot (\lambda_{11}(\mathbf{G}))^{-n}} \\ &= \frac{\eta}{2} \cdot \frac{\tau_i}{\sigma_1 - \sigma_i} + o(\eta) \end{aligned}$$

由 命题 5.3 可知

$$\lim_{n\to\infty} \bar{v}_n(\boldsymbol{\phi}_0) = \left(1 + \lim_{n\to\infty} \sum_{i=2}^{d} \frac{\pi_n^{(i)}(\boldsymbol{\phi}_0)}{\pi_n^{(1)}(\boldsymbol{\phi}_0)} \right)^{-1}$$

$$= \left(1 + \frac{\eta}{2} \cdot \sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} + o(\eta)\right)^{-1}$$
$$= 1 - \frac{\eta}{2} \cdot \sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} + o(\eta),$$

于是有

$$\bar{v}_\infty^{\mathsf{c}} \triangleq \lim_{n\to\infty} \bar{v}_n^{\mathsf{c}}(\boldsymbol{\phi}_0) = \frac{\eta}{2} \cdot \sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} + o(\eta).$$

同理可得关于 $\bar{\rho}_n^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 的结论。

最后，由 (5-29) 知收敛率为 $\mathbf{G}$ 第二大特征值与最大特征值的比值，亦即 $\gamma_{12}(\eta)$。根据 (5-32) 可求得 $r = \gamma_{12}(\eta) = \lambda_{12}(\mathbf{G})/\lambda_{11}(\mathbf{G}) = 1 - \eta(\sigma_1 - \sigma_2) + o(\eta)$。 □

由 定理 5.4 可得类似于第 5.2 节中的折中关系：

$$\bar{v}_\infty^{\mathsf{c}} = r^{\mathsf{c}} \cdot \frac{1}{2(\sigma_1 - \sigma_2)} \sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_j} + o(r^{\mathsf{c}}),$$
$$\bar{\rho}_\infty^{\mathsf{c}} = r^{\mathsf{c}} \cdot \frac{1}{2(\sigma_1 - \sigma_2)} \sum_{i=2}^{d} \tau_i + o(r^{\mathsf{c}}),$$

其中 $r^{\mathsf{c}} \triangleq 1 - r$。

虽然该折中关系依赖于 η 恒定的假设，但其可提供对机器学习实践中所采用的复杂学习率方案的理论解释。例如在阶跃衰减方案 (Step Decay Schedule[64]) 中，学习率将每隔常数期 (Epoch) 以固定比例衰减。若初始学习率较小，可利用所建立折中关系将该方案解释为在训练初期加速算法收敛，而在训练末期通过衰减学习率获得较小的泛化误差。一般情况下对该方案泛化误差上界的分析可参考 [65]。对比一般情况下误差界的分析工作，我们在小学习率机制下给出了泛化误差取值的精确表达式，从而可有效指导实践中学习率方案的设计。

由于实践中样本数 N 通常为有限值，分析有限的 N 次迭代后的泛化误差将更具有实际意义。为叙述大样本小学习率机制下泛化误差的结论，首先介绍该机制中的小量定义如下。

定义 5.1： 给定 η 与 N 的函数 f，令 $f(\eta, N) = \hat{o}(1)$ 表示存在函数 $g(t)$ 及 $h(t)$，使得

$$\lim_{t\to 0^+} g(t) = \lim_{t\to+\infty} h(t) = 0,$$

且对任意 $\eta > 0$ 及 $N > 0$ 都有

$$|f(\eta, N)| \leqslant |g(\eta)| + |h(N\eta)|.$$

关于一般情况下的泛化误差可得如下定理，其证明可参见 附录 C.5。

定理 5.5： 在大样本小学习率分析机制下，N 轮迭代之后的泛化误差可表示为

$$\bar{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0) = \hat{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)(1 + \hat{o}(1)) \tag{5-36}$$

$$\bar{\rho}_N^{\mathsf{C}}(\boldsymbol{\phi}_0) = \hat{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)(1 + \hat{o}(1)), \tag{5-37}$$

其中 $\hat{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 及 $\hat{\rho}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 定义为

$$\hat{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0) \triangleq e^{2(\sigma_2-\sigma_1)N\eta} \cdot \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0 \rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2} + \frac{\eta}{2} \sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} \tag{5-38}$$

$$\hat{\rho}_N^{\mathsf{C}}(\boldsymbol{\phi}_0) \triangleq (\sigma_1 - \sigma_2) e^{2(\sigma_2-\sigma_1)N\eta} \cdot \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0 \rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2} + \frac{\eta}{2} \sum_{i=2}^{d} \tau_i. \tag{5-39}$$

由定理 5.5 可知，在大样本小学习率分析机制下，算法接近收敛且最终泛化误差值很小。其次，$\bar{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 与 $\bar{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 均可解释为无噪条件下的泛化误差与渐进泛化误差的和。再次，泛化误差理论值 $\hat{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 及 $\hat{\rho}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 关于学习率 η 变化的图像均呈 U 型曲线，如图 5.2所示。由此，我们的分析为特征问题提供了训练误差受学习率大小影响的典型关系[6] 的理论解释，可视为对凸问题中类似结论的推广[24,26]。

此外，易知最小化 $\hat{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 的最优学习率为

$$\eta_v^* = \frac{\log(C_v N)}{2(\sigma_1 - \sigma_2)N}, \tag{5-40}$$

其中

$$C_v \triangleq 4(\sigma_1 - \sigma_2) \cdot \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0 \rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2} \cdot \left(\sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} \right)^{-1}.$$

类似可得最小化 $\hat{\rho}_N^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 的最优学习率

$$\eta_\rho^* = \frac{\log(C_\rho N)}{2(\sigma_1 - \sigma_2)N} \tag{5-41}$$

其中

$$C_\rho \triangleq 4(\sigma_1 - \sigma_2)^2 \cdot \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0 \rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0 \rangle^2} \cdot \left(\sum_{i=2}^{d} \tau_i \right)^{-1}.$$

注释 5.3： 注意到 η_v^* 与 η_ρ^* 表达式均为 $\frac{\log(CN)}{2(\sigma_1-\sigma_2)N}$ 的形式，其中常数 C 分别取 C_v 及 C_ρ。事实上，可验证 C 取值不影响泛化误差的渐进行为。因此可不妨令 $C=1$ 并将 $\eta=\frac{\log N}{2(\sigma_1-\sigma_2)N}$ 作为最优学习率。

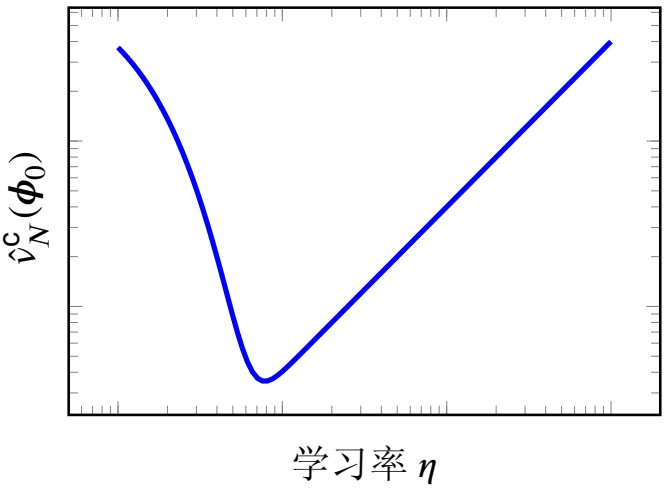


图 5.2　泛化误差理论值 $\hat{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 及学习率 η 的关系，其中横纵坐标均取对数。泛化误差 $\hat{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 关于学习率有类似的变化趋势形。

与定理 5.3 结果类似，可对泛化误差作如下概率层面的刻画，其证明可参见附录 C.6。

定理 5.6： 给定初始向量 $\boldsymbol{\phi}_0$ 及充分大的 N，若学习率取 $\eta=\frac{\log N}{2(\sigma_1-\sigma_2)N}$, 则对任意 $\delta\in(0,1)$ 有

$$\mathbb{P}\left(\nu^{\mathsf{c}}(\boldsymbol{\phi}_N)<\frac{C_0}{\delta}\cdot\frac{\log^2 N}{N}\middle|\boldsymbol{\phi}_0\right)>1-\delta \tag{5-42}$$

$$\mathbb{P}\left(\rho^{\mathsf{c}}(\boldsymbol{\phi}_N)<\frac{\sigma_1 C_0}{\delta}\cdot\frac{\log^2 N}{N}\middle|\boldsymbol{\phi}_0\right)>1-\delta, \tag{5-43}$$

其中 $C_0\triangleq\frac{\tau_1}{2(\sigma_1-\sigma_2)^2}$。

注释 5.4： 定理 5.6 给出了 $\Theta\left(\frac{\log^2 N}{N}\right)$ 的误差，与信息论上的极小极大界 $1/N$[66] 只差一个对数多项式因子。尽管 $1/N$ 界的推导依赖于 $\mathbf{x}_n$ 服从 sub-Gaussian 分布的假设，这里的分析只需假设 $\mathbf{Z}_n$ 具有有限的二阶矩，或假设流式主成分分析中的 $\mathbf{x}_n$ 具有有限的四阶矩。

5.3.3　小批量训练的 Oja 算法

在实际训练学习算法时，参数更新通常基于一小批样本进行，对应训练机制称为小批量训练[6]。由于在小批量样本上的计算可并行化，采用小批量训练可以有效地提高计算效率；对于合适的小批量大小取值，小批量学习还可以降低泛化误差[24,67]。本节将基于 Oja 算法的特例，考察小批量学习对泛化误差的影响。具体地，设总迭代次数 N 给定，对应于总样本数 (参见例 5.1)。除此之外，使用 m 表示小批量大小并假设 $N = km$，则基于小批量训练的 Oja 算法流程可概括为 算法 4，其中 $\mathbf{A}\boldsymbol{\phi}_{n-1}$ 在一个小批量中 m 次估计的平均将用于计算 $\boldsymbol{\phi}_n$。

算法 4 小批量训练的 Oja 算法

输入: $\boldsymbol{\phi}_0$.
for $n = 1$ **to** k **do**
　$\boldsymbol{\psi}_n \leftarrow \mathbf{0}$
　for $i = 1$ **to** m **do**
　　$\boldsymbol{\psi}_n \leftarrow \boldsymbol{\psi}_n + \mathbf{A}\boldsymbol{\phi}_{n-1} + \boldsymbol{\zeta}_{(n-1)m+i}$
　end for
　$\boldsymbol{\psi}_n \leftarrow \boldsymbol{\psi}_n/m$
　$\boldsymbol{\phi}_n \leftarrow \boldsymbol{\phi}_{n-1} + \eta\boldsymbol{\psi}_n$
end for
输出: $\boldsymbol{\phi}_k/\|\boldsymbol{\phi}_k\|$.

注意到算法 3 可视为小批量大小取 $m = 1$ 时小批量训练的特例。与 算法 3 相比，小批量训练会对泛化误差产生正反两种效用：一方面，在小批量中对噪声求平均将使得每次更新更为准确；另一方面，更新次数也将减少为原来的 $1/m$。为量化小批量训练的影响，记 $\bar{v}_n^{\mathsf{c}}\left(\boldsymbol{\phi}_0;\eta,m\right)$ 及 $\bar{\rho}_n^{\mathsf{c}}\left(\boldsymbol{\phi}_0;\eta,m\right)$ 为小批量训练中，$\boldsymbol{\phi}_n$ 所给出的平均泛化误差 [参见(5-23)]，其中 η 表示相应的学习率。以下结果表明，在泛化误差方面，使用小批量大小为 m 的小批量训练等价于将学习率减小为原先的 $1/m$。

定理 5.7： 在大样本及小学习率机制下，设批量大小 m 为给定常数，则有

$$\bar{v}_k^{\mathsf{c}}(\boldsymbol{\phi}_0;\eta,m) = \bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0;\eta_{\mathsf{eq}},1)(1+\hat{o}(1)), \tag{5-44}$$

$$\bar{\rho}_k^{\mathsf{c}}(\boldsymbol{\phi}_0;\eta,m) = \bar{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0;\eta_{\mathsf{eq}},1)(1+\hat{o}(1)), \tag{5-45}$$

其中 等价学习率 η_{eq} 定义为

$$\eta_{\mathsf{eq}} \triangleq \frac{\eta}{m}, \tag{5-46}$$

且 $\hat{o}(1)$ 定义由定义 5.1 给出。

证明 首先注意到由算法 4 中的更新步骤有

$$\boldsymbol{\phi}_n = (\mathbf{I} + \eta\mathbf{A})\boldsymbol{\phi}_{n-1} + \eta\hat{\boldsymbol{\xi}}_n, \quad n = 1, \cdots, k, \tag{5-47}$$

其中

$$\hat{\boldsymbol{\xi}}_n \triangleq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\zeta}_{(n-1)m+i} = \hat{\mathbf{Z}}_n\boldsymbol{\phi}_{n-1}$$

且

$$\hat{\mathbf{Z}}_n \triangleq \frac{1}{m}\sum_{i=1}^{m} \mathbf{Z}_{(n-1)m+i}. \tag{5-48}$$

因此 (5-47) 等价于 Oja 算法的 k 次迭代，且由 (5-28) 及 (5-48) 可知其对应的参数 τ 为 $\hat{\tau}_i = \tau_i/m$ $(i = 1, \cdots, d)$。于是，由定理 5.5 得

$$\begin{aligned}
\bar{v}_k^{\mathsf{C}}(\boldsymbol{\phi}_0; \eta, m) &= \left[e^{2(\sigma_2-\sigma_1)k\eta} \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0\rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0\rangle^2} + \frac{\eta}{2}\sum_{i=2}^{d} \frac{\hat{\tau}_i}{\sigma_1 - \sigma_i} \right] \cdot (1 + \hat{o}(1)), \\
&= \left[e^{2(\sigma_2-\sigma_1)N\eta_{\mathsf{eq}}} \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0\rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0\rangle^2} + \frac{\eta_{\mathsf{eq}}}{2}\sum_{i=2}^{d} \frac{\tau_i}{\sigma_1 - \sigma_i} \right] \cdot (1 + \hat{o}(1)), \\
&= \bar{v}_N^{\mathsf{C}}(\boldsymbol{\phi}_0; \eta_{\mathsf{eq}}, 1) \cdot (1 + \hat{o}(1)).
\end{aligned}$$

同理可得 (5-45)。 □

因此，使用批量大小为 m 学习率为 η 的小批量训练等价于将学习率 η 衰减为 $\eta_{\mathsf{eq}} = \eta/m$，从而将学习率 η 与批量大小 m 等比例放大将不影响泛化误差的大小。在实际学习问题的应用中，由于小批量训练可并行操作，使用该等比例放大操作可得到正比于批量大小的加速比，且不会使泛化误差恶化。该性质也称作线性缩放准则，其在加速深度神经网络训练方面有广泛的应用[68-70]。

5.4　实验结果

本节在仿真数据及 MNIST 数据集[10] 上开展了一系列实验，以检验第 5.3 节中的理论结果。

5.4.1　仿真数据

在该实验中，取 $d = 10$ 并随机生成相应的半正定矩阵 $\mathbf{A}$。在此基础上，根据(5-20)生成 Oja 算法中的噪声，其中诸 $\mathbf{Z}_n$ 为独立同分布的零均值 Gaussian 矩阵。接着，在随机生成初始向量 $\boldsymbol{\phi}_0$ 后，将算法 3 重复执行 20 次，每次执行包含 $N = 10^6$ 次迭代。于是平均泛化误差可通过 (5-24) 计算，其中重复试验所得泛化误差的经验平均可作为数学期望的估计。

首先考察不同学习率 η 设定下的泛化误差。为此，根据算法 3 重复运行 20 次的结果计算，可得到不同学习率下的 $\bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 以及 $\bar{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)$，如图 5.3 所示，其中实线对应由定理 5.5 给出的理论结果。由图可知，仿真结果与理论结果高度一致，尤其是在学习率 η 相对较小的情况下。

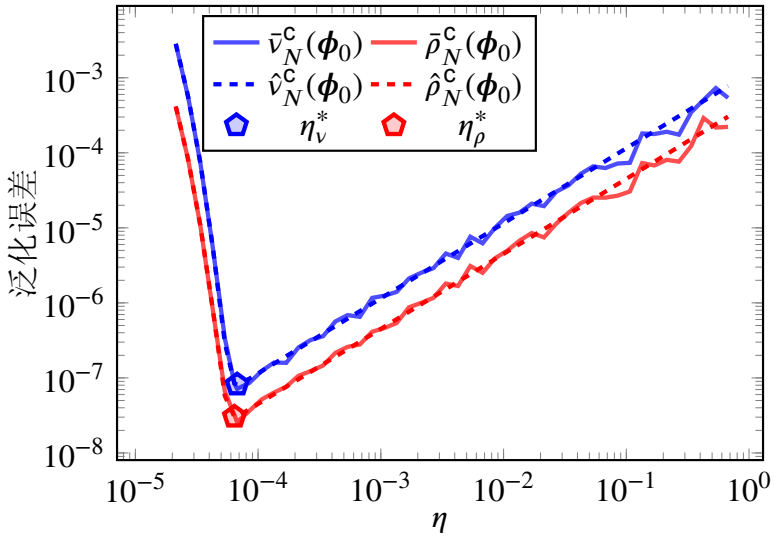


图 5.3　泛化误差随学习率的变化关系，其中实线和虚线分别表示仿真所得泛化误差及其理论值。

其次考察泛化误差在迭代过程中的收敛情况。具体地，分别取学习率 $\eta = 2^i\eta_v^*, i = -1, 0, 1, 2, 3$，其中 η_v^* 定义由 (5-40) 给出，并考察 $\bar{v}_n^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 随 n 变化的趋势。相应结果如图 5.4 所示，其中泛化误差的理论值 $\hat{v}_n^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 可根据 (5-38) 计算得到。该结果表明再次验证了理论结果的有效性。除此之外，图 5.4 也直观地表明了泛化误差 $\bar{v}_n^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 及收敛率的折中关系是如何受学习率选择影响的。当选用 Rayleigh 商 $\bar{\rho}_n^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 作为泛化误差度量时，亦可得到类似结果。

再次，考察小批量训练中的泛化误差。特别的，令批量大小分别取 $m = 1, 8, 32$。图 5.5 给出了泛化误差 $\bar{v}_{N/m}^{\mathsf{c}}(\boldsymbol{\phi}_0; \eta, m)$ 及 $\bar{\rho}_{N/m}^{\mathsf{c}}(\boldsymbol{\phi}_0; \eta, m)$ 关于等价学习率 $\eta_{\mathsf{eq}} = \eta/m$ 变化的曲线，相应结果验证了第 5.3.3 节中关于批量大小的讨论。

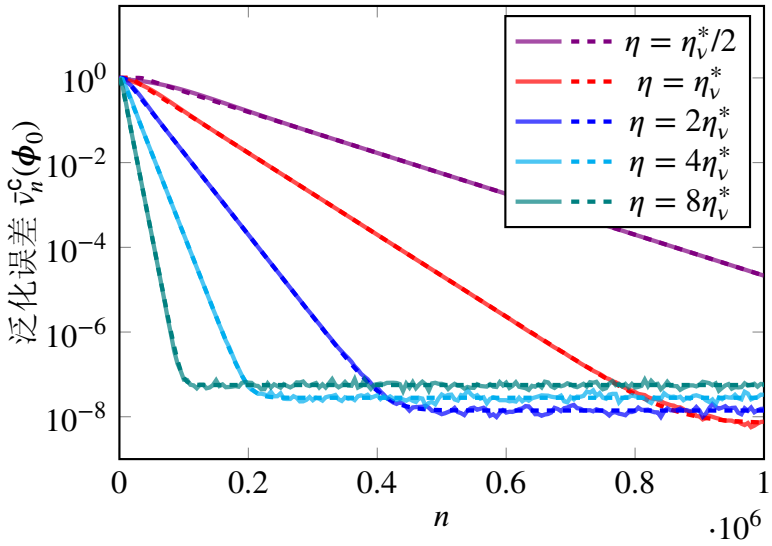


图 5.4　在不同学习率 η 下，泛化误差 $\bar{v}_n^{\mathsf{C}}(\boldsymbol{\phi}_0)$ 随迭代次数 n 的变化，其中实线和虚线分别表示仿真所得泛化误差及相应的理论值 $\hat{v}_n(\boldsymbol{\phi}_0)$。

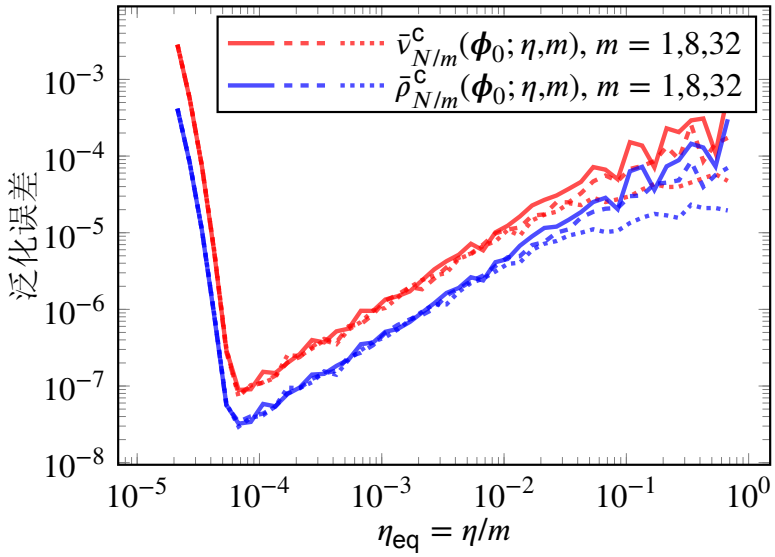


图 5.5　在小批量大小 m 取不同值时，泛化误差随等价学习率 $\eta_{\mathsf{eq}} = \eta/m$ 的变化关系。

5.4.2　MNIST 手写体数据集

进一步地，我们在 MNIST 手写体数据集[10] 上验证理论结果。该数据集包含 60,000 个训练样本及 10,000 个测试样本。首先，利用 算法 4 计算训练集的主成分。其次，计算测试样本的协方差矩阵，作为 $\mathbf{A}$ 的真实值，并由此计算得主成分 $\mathbf{v}_1, \cdots, \mathbf{v}_d$，用于计算泛化误差。注意到由于 MNIST 数据集对应的参数 τ_i 的取值无法准确地从样本中估计，很难精确地给出其理论性能用于实验对比。为此，我们转而在 MNIST 数据集中对小批量训练的线性缩放准则进行验证。具体地，对于某个给定的输入 $\boldsymbol{\phi}_0$，令小批量大小分别取 $m = 64, 128, 256$。对每个 m 的取值，在 $N = 50,000$ 个训练样本上重复运行 Oja 算法 20 次，并且在每次运行之前将样本

重新随机排列。实验所得泛化误差 $\bar{v}^{\mathsf{c}}_{N/m}(\boldsymbol{\phi}_0;\eta,m)$ 及 $\bar{\rho}^{\mathsf{c}}_{N/m}(\boldsymbol{\phi}_0;\eta,m)$ 关于等价学习率 $\eta_{\mathsf{eq}}=\eta/m$ 的变化趋势如图 5.6 所示，相应结果也与第 5.3.3 节中的理论分析相符。

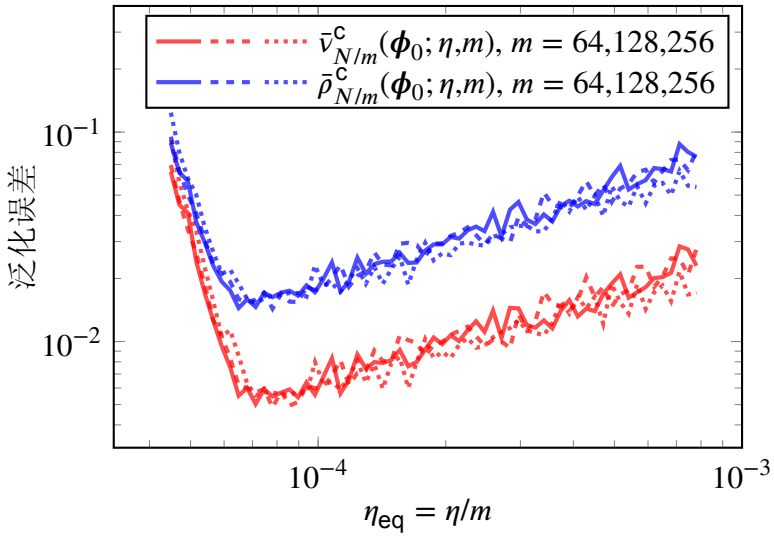


图 5.6　MNIST 数据集上泛化误差随等价学习率 $\eta_{\mathsf{eq}}=\eta/m$ 的变化关系，其中小批量大小分别取 $m=64,128,256$。

5.5　本章小结

由深度神经网络实际训练过程的梯度下降法出发，本章考察训练计算过程中泛化误差与计算效率的最优折中。为此，首先考虑局部分析机制下的随机梯度下降过程，将其简化为鲁棒交替条件期望算法，并分析了鲁棒条件期望算法中计算效率与泛化误差满足的折中关系。在此基础上，进一步将结论推广到求解一般特征问题的 Oja 算法，建立了泛化误差的解析表达并给出最优学习率的选择。此外，对小批量训练的 Oja 算法的分析表明，小批量大小对泛化误差的影响等价于改变学习率。结合深度神经网络的奇异值分解数学结构，本章的结论可为实际深度神经网络中更为复杂的超参数机制设计提供理论指导。

第 6 章　机器学习算法设计

6.1　本章引言

基于局部信息几何的分析工具，本章将信息论用于指导机器学习算法设计，特别是在深度学习框架下的算法实现。具体而言，算法设计范式可概括为以下三步：首先，基于信息论中经典的信息度量的分析，构建相应概率空间的优化问题；其次，基于局部信息几何分析中概率分布与信息向量的对应关系，将前述优化问题转化为有限维空间中信息向量的优化问题，如特征值分解或奇异值分解问题等；最后，基于信息向量与数据特征的对应关系，将信息向量的优化问题进一步转化为函数空间内特征的优化问题，并构造深度神经网络求解该最优特征。在该设计范式下，本章给出了基于深度学习从数据样本计算最大相关函数的框架，并特别考察了其在有监督学习问题中的应用。类似地，基于信息论中经典的总相关 (Total Correlation) 度量，本章设计了高效的无监督特征提取算法。对算法的分析表明，由该设计范式给出的算法可视为对主成分分析、线性判别分析等经典机器学习算法的推广，从而具有理论性能保障；在一系列常见数据集上实验结果进一步检验了所设计算法的有效性。

本章具体内容安排如下：第 6.2 节简单介绍了最大相关函数计算的神经网络实现；第 6.3 节考虑最大相关函数在有监督学习问题中的应用，提出了最大相关回归方法并对其性质进行了分析；第 6.4 节考虑基于信息论中的总相关的最优无监督特征提取问题，考察了最优特征的理论性质并设计了基于深度学习的计算框架；最后，第 6.5 节介绍了所设计算法在常见数据集上的结果，第 6.6 节总结了全章内容。

6.2　最大相关函数学习

给定随机变量 $\boldsymbol{X}, \boldsymbol{Y}$，基于第 3 章中讨论可知，最大化 定义 3.4 中的 H 评分函数 $H(\mathbf{f}(\boldsymbol{X}), \mathbf{g}(\boldsymbol{Y}))$ 等价于求解典型相关矩阵 $\tilde{\mathbf{B}}$ 的奇异向量，其最优解为 定义 2.4 所给出的最大相关函数。由于 H 评分函数可高效地从训练样本中计算 (计算细节可参考 [71] 的算法 1)，故该性质自然给出了从数据样本中学习最大相关函数的框架，如图 6.1 所示。其中用于特征提取的 $\mathrm{NN}_{\mathbf{f}}$ 与 $\mathrm{NN}_{\mathbf{g}}$ 可选用与输入数据对应的神经网络结构，例如，对图像输入可用卷积神经网络提取特征。

注意到 H 评分函数为经典神经网络的损失函数的局部近似，因此该学习框架可作为局部信息几何方法用于指导机器学习算法设计的实例之一。后面的实验结

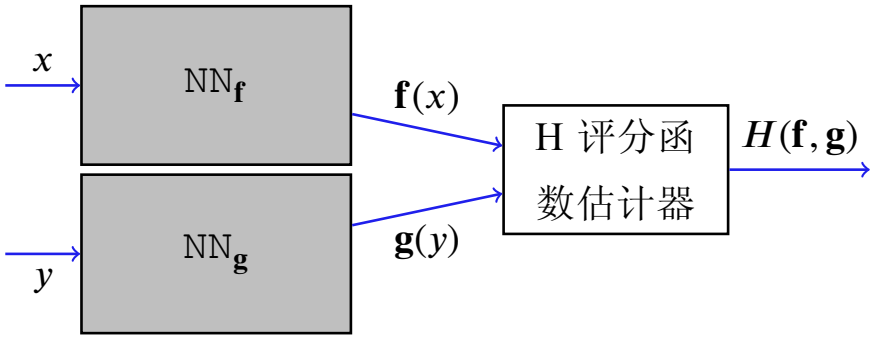


图 6.1　最大相关函数学习的神经网络架构，其中 $\mathrm{NN}_{\mathbf{f}}$ 用于从输入 x 中生成特征 $\mathbf{f}(x)$, $\mathrm{NN}_{\mathbf{g}}$ 用于从输入 x 中生成特征 $\mathbf{f}(x)$。

果中可以看出，该框架所给出的最大相关函数相比已有启发式设计方法可取得明显的性能优势。对该框架更深入的讨论可参考 [71]，其由典型相关矩阵 $\tilde{\mathbf{B}}$ 的低秩恢复问题出发引入该学习框架，并介绍了其在推荐系统等实际系统中的应用。

6.3　有监督学习：最大相关回归

本节讨论当 Y 为离散标签时，X 与 Y 的最大相关函数计算，并将其用于预测标签 Y 的有监督学习问题中。此时可证明，最大相关函数将自然导出后验概率 $P_{Y|X}$ 的估计，因此该方法可视为对后验概率的回归问题，称为最大相关回归[72]。

6.3.1　问题构建

最大相关回归的关键思路是在函数族中 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}$ 寻找分布近似 $P_{Y|X}$，其中 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}$ 定义如下。

定义 6.1： 分别给定 X 和 Y 的 k 维函数 $\mathbf{f}(\cdot)$ 及 $\mathbf{g}(\cdot)$，及 Y 的标量函数 $b(\cdot)$，定义 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}$ 为

$$P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x) \triangleq P_Y(y)\left(1+\mathbf{f}^{\mathsf{T}}(x)\mathbf{g}(y)+b(y)\right). \tag{6-1}$$

注意 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)$ 可视为 Softmax 函数 (3-2) 的局部展开。由于 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)$ 可能取负值，不能用传统的交叉熵作为 $P_{Y|X}$ 及 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}$ 差距的度量。这里采用 χ^2 散度[①]

$$L(\mathbf{f},\mathbf{g},b) \triangleq \sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} \frac{\left[P_{Y|X}(y|x)-P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)\right]^2}{P_Y(y)} \tag{6-2}$$

① 为便于表述，这里仍采用离散 X 对应的求和符号。对连续的变量 X，只需将对 x 的求和改为相应的积分。

作为 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}$ 逼近 $P_{Y|X}$ 程度的度量，其操作意义将在后文解释最大相关回归与 HGR 最大相关及神经网络的联系时进一步明确。易知 $L(\mathbf{f},\mathbf{g},b)\geqslant 0$，其中等号成立当且仅当对所有 x,y 都有 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)=P_{Y|X}(y|x)$。下面给出最大相关回归的正式定义。

定义 6.2 (最大相关回归)： 给定输入特征 $\mathbf{f}(X)\in\mathbb{R}^k$，最大相关回归 (Maximal Correlation Regression, MCR) 使用含参分布 $P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}$ 作为后验概率 $P_{Y|X}$ 的建模，并选取最优的 $\mathbf{g}\colon\mathcal{Y}\mapsto\mathbb{R}^k$ 及 $b\colon\mathcal{Y}\mapsto\mathbb{R}$ 以最小化 (6-2) 中的 $L(\mathbf{f},\mathbf{g},b)$。

不同于 Softmax 回归，最大相关回归中的最优参数 $\mathbf{g}^*$ 及 b^* 有解析表达式，如下所述。

命题 6.1： 对给定特征 $\mathbf{f}(X)$ 及任意 $y\in\mathcal{Y}$，有[①]

$$\mathbf{g}^*(y)=\boldsymbol{\Lambda}_{\mathbf{f}}^{-1}\mathbb{E}\left[\tilde{\mathbf{f}}(X)\middle|Y=y\right] \tag{6-3a}$$

$$b^*(y)=-\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}^*(y). \tag{6-3b}$$

证明 参见附录 D.1。 □

注意 (6-3) 中的协方差矩阵 $\boldsymbol{\Lambda}_{\mathbf{f}}$、均值 $\boldsymbol{\mu}_{\mathbf{f}}$ 以及条件期望 $\mathbb{E}\left[\tilde{\mathbf{f}}(X)\middle|Y=y\right]$ 都可以直接从数据样本中估计得到，从而最优参数 $\mathbf{g}^*$ 及 b^* 可从训练样本中高效求得，具体计算过程可总结为算法 5。

对于新的数据样本 x，其标签 y 可通过 最大后验概率 (Maximum a Posteriori, MAP) 判决准则进行预测，并得

$$\hat{y}(x)=\arg\max_{y'\in\mathcal{Y}}\ P_{Y|X}^{(\mathbf{f},\mathbf{g}^*,b^*)}(y'|x) \tag{6-4}$$

作为最大相关回归的预测结果。

6.3.2　基于深度学习框架的最大相关回归

在第 6.3.1 节的讨论中假定回归问题中的特征 $\mathbf{f}$ 预先给定。一般情况下，$\mathbf{f}$ 也可由参数为 $\boldsymbol{\theta}$ 的模型 $\mathbf{f}_{\boldsymbol{\theta}}\triangleq\mathbf{f}(\cdot;\boldsymbol{\theta})$ 生成。例如，对深度神经网络 $\boldsymbol{\theta}$ 对应于所有隐层的权重与偏置项。该情况下，损失函数 $L(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g},b)$ 应对 $\mathbf{g},b$ 及 $\boldsymbol{\theta}$ 联合优化。具体而言，可基于随机梯度下降及反向传播算法[73] 利用训练样本对 $\mathbf{g},b$ 及 $\boldsymbol{\theta}$ 优化。注意由于 $P_{Y|X}(y|x)$ 未知，$L(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g},b)$ 的值无法直接从训练样本中估计。但以下结果表明，对 $L(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g},b)$ 的优化可转化为对 H 评分函数优化的问题，从而可基于样本求解。

① 与第 3 章类似，本章的分析中，采用 “~” 符号表示减去均值后的函数，且使用 $\boldsymbol{\Lambda}$ 表示协方差矩阵。例如：如 $\tilde{\mathbf{f}}(x)\triangleq\mathbf{f}(x)-\mathbb{E}[\mathbf{f}(X)],x\in\mathcal{X}$；$\boldsymbol{\Lambda}_{\mathbf{f}}$ 为 $\mathbf{f}$ 的协方差矩阵。

算法 5 最大相关回归中的参数估计

输入: 特征及标签的 n 个样本 $\{(\mathbf{f}(x_i), y_i)\}_{i=1}^n$.

计算均值，归一化得零均值特征:

$$\hat{\boldsymbol{\mu}}_{\mathbf{f}} \leftarrow \frac{1}{n}\sum_{i=1}^n \mathbf{f}(x_i)$$
$$\mathbf{f}(x_i) \leftarrow \mathbf{f}(x_i) - \hat{\boldsymbol{\mu}}_{\mathbf{f}}, \quad i = 1, \cdots, n$$

计算输入特征的协方差矩阵:

$$\hat{\boldsymbol{\Lambda}}_{\mathbf{f}} \leftarrow \frac{1}{n}\sum_{i=1}^n \mathbf{f}(x_i)\mathbf{f}^{\mathsf{T}}(x_i)$$

for $y = 1$ **to** $|\mathcal{Y}|$ **do**

$$n_y \leftarrow \sum_{i=1}^n \mathbb{1}_{\{y=y_i\}}$$
$$\hat{\boldsymbol{\mu}}_{\mathbf{f}|Y=y} \leftarrow \frac{1}{n_y}\sum_{i=1}^n \mathbf{f}(x_i)\cdot\mathbb{1}_{\{y=y_i\}}$$

end for

for $y = 1$ **to** $|\mathcal{Y}|$ **do**

$$\mathbf{g}^*(y) \leftarrow \hat{\boldsymbol{\Lambda}}_{\mathbf{f}}^{-1}\hat{\boldsymbol{\mu}}_{\mathbf{f}|Y=y}$$
$$b^*(y) \leftarrow -\hat{\boldsymbol{\mu}}_{\mathbf{f}}\mathbf{g}(y)$$

end for

输出: 对所有 $y \in \mathcal{Y}$，$\mathbf{g}^*(y), b^*(y)$ 的取值。

命题 6.2: 令 $(\boldsymbol{\theta}^*, \mathbf{g}^*, b^*)$ 表示使 $L(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{g}, b)$ 最小化的最优参数，则有 $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{\mathrm{H}}$ 以及，对任意 $y \in \mathcal{Y}$,

$$\mathbf{g}^*(y) = \tilde{\mathbf{g}}_{\mathrm{H}}(y), \quad b^*(y) = -\boldsymbol{\mu}_{\mathbf{f}^*}^{\mathsf{T}}\mathbf{g}^*(y), \tag{6-5}$$

其中 $\boldsymbol{\theta}_{\mathrm{H}}$ 与 $\mathbf{g}_{\mathrm{H}}$ 为使得 H 评分函数

$$H(\mathbf{f}_\theta, \mathbf{g}) \triangleq \mathbb{E}\left[\tilde{\mathbf{f}}^{\mathsf{T}}(X;\boldsymbol{\theta})\tilde{\mathbf{g}}(Y)\right] - \frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Lambda}_{\mathbf{f}_\theta}\boldsymbol{\Lambda}_{\mathbf{g}}\right\}, \tag{6-6}$$

最大化的参数，且 $\boldsymbol{\mu}_{\mathbf{f}^*} \triangleq \mathbb{E}[\mathbf{f}(X;\boldsymbol{\theta}^*)]$。

证明 参见附录 D.2. □

注意到 H 评分函数可基于数据样本高效估计，具体细节可参考 [71] 的算法 1，因此可使用负的 H 评分函数 $-H(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{g})$ 作为损失函数，基于随机梯度下降法训练最优参数 $\boldsymbol{\theta}_{\mathrm{H}}$ 及 $\mathbf{g}_{\mathrm{H}}$。该训练所需的神经网络框架的讨论可参考附录 D.3。

在训练得到 $\boldsymbol{\theta}_{\mathrm{H}}$ 与 $\mathbf{g}_{\mathrm{H}}$ 后，基于命题 6.2 可从数据样本中恢复参数 $\boldsymbol{\theta}^*, \mathbf{g}^*$ 及 b^* 的值，相应计算过程可总结为算法 6。

对于新观测到的 x，由最大后验概率准则预测得到的标签 $\hat{y}(x)$ 为

$$\hat{y}(x) = \arg\max_{y' \in \mathcal{Y}} P_{Y|X}^{(\mathbf{f}_{\boldsymbol{\theta}^*}, \mathbf{g}^*, b^*)}(y'|x). \tag{6-7}$$

算法 6 最大相关回归中的参数恢复

输入: 数据样本 $\{x_i\}_{i=1}^n$, $\boldsymbol{\theta}_{\mathrm{H}}, \mathbf{g}_{\mathrm{H}}$，函数 $\mathbf{f}_{\boldsymbol{\theta}}$

$\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_{\mathrm{H}}$

计算均值:

$$\hat{\boldsymbol{\mu}}_{\mathbf{f}} \leftarrow \frac{1}{m}\sum_{i=1}^n \mathbf{f}(x_i;\boldsymbol{\theta}^*)$$

$$\hat{\boldsymbol{\mu}}_{\mathbf{g}} \leftarrow \sum_{y\in\mathcal{Y}} P_Y(y)\mathbf{g}_{\mathrm{H}}(y)$$

for $y = 1$ **to** $|\mathcal{Y}|$ **do**

　$\mathbf{g}^*(y) \leftarrow \mathbf{g}_{\mathrm{H}}(y) - \hat{\boldsymbol{\mu}}_{\mathbf{g}}$

　$b^*(y) \leftarrow -\hat{\boldsymbol{\mu}}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}^*(y)$

end for

输出: $\boldsymbol{\theta}^*$ 及 $\mathbf{g}^*(y), b^*(y)$ 对所有 $y \in \mathcal{Y}$ 的取值

6.3.3　理论性质

首先，可证明优化最大相关回归的参数等价于求解典型相关矩阵的低秩恢复问题。为便于表述，设 X 为离散随机变量，并令 $\boldsymbol{\Xi}^Y$ 及 $\boldsymbol{\Xi}^X$ 为由 定义 2.1 给出的函数 $\tilde{\mathbf{f}}, \tilde{\mathbf{g}}$ 所对应的等价矩阵表示：

$$\boldsymbol{\Xi}^X = \left[\tilde{\mathbf{f}}(1)\sqrt{P_X(1)}, \cdots, \tilde{\mathbf{f}}(|\mathcal{X}|)\sqrt{P_X(|\mathcal{X}|)}\right]^{\mathsf{T}}, \tag{6-8a}$$

$$\boldsymbol{\Xi}^Y = \left[\tilde{\mathbf{g}}(1)\sqrt{P_Y(1)}, \cdots, \tilde{\mathbf{g}}(|\mathcal{Y}|)\sqrt{P_Y(|\mathcal{Y}|)}\right]^{\mathsf{T}}. \tag{6-8b}$$

下述命题给出了最大相关回归问题的低秩恢复结构。

命题 6.3： 给定 $\mathbf{f}$、$\mathbf{g}$ 及 b，可得

$$L(\mathbf{f},\mathbf{g},b) = \left\|\tilde{\mathbf{B}} - \boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2 + \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}} + \sum_{y\in\mathcal{Y}} P_Y(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right]^2, \tag{6-9}$$

其中 $\tilde{\mathbf{B}} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ 为 X 与 Y 的典型相关矩阵 (可参考定义 2.3)。

证明 参见附录 D.4。 □

注意到 $\boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}} + \sum_{y\in\mathcal{Y}} P_Y(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right]^2 \geqslant 0$ 且等号成立当且仅当

$$\boldsymbol{\mu}_{\mathbf{g}} = \mathbf{0} \quad \text{且} \quad b(y) = -\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y),\ \forall\, y \in \mathcal{Y}. \tag{6-10}$$

故为使损失函数 $L(\mathbf{f},\mathbf{g},b)$ 最小化，应令 $\boldsymbol{\mu}_{\mathbf{g}}$ 及 b 取 (6-10) 中的值。于是最大相关回归可化简为最小化 $\left\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}$ 的问题，即使用秩 k 矩阵 $\left[\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\right]$ 逼近典型相关矩阵 $\tilde{\mathbf{B}}$。由 Eckart–Young–Mirsky 定理[56] 知最优的 $\boldsymbol{\Xi}^{Y}$ 与 $\boldsymbol{\Xi}^{X}$ 将分别对应 $\tilde{\mathbf{B}}$ 前 k 个左奇异向量和右奇异向量。该对应关系可用于建立最大相关回归与 HGR 最大相关问题的联系。

注意到命题 6.1 中 $\mathbf{g}^{*}$ 的计算可解释为交替条件期望算法中的一步 [参考 (3-11)]。此外，对给定的 $\mathbf{f}$，当 $\mathbf{g}$ 与 b 均取最优值 (6-3) 时，相应的损失函数 $L(\mathbf{f},\mathbf{g}^{*},b^{*})$ 可表示成

$$L(\mathbf{f},\mathbf{g}^{*},b^{*})=\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^{2}-2H_{Y}(\mathbf{f}),$$

其中 $H_{Y}(\mathbf{f})$ 为 $\mathbf{f}(X)$ 的单边 H 评分函数，其定义为 (参考定义 3.4)

$$H_{Y}(\mathbf{f})\triangleq\mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{\mathbf{f}}^{-1/2}\mathbb{E}\left[\tilde{\mathbf{f}}(X)\middle|Y\right]\right\|^{2}\right]. \tag{6-11}$$

以下定理表明，单边 H 评分函数可用于刻画最大相关回归的预测准确度。

定理 6.1： 对给定 $\mathbf{f}$，令 $\mathrm{ACC}(\mathbf{f})$ 记为由 (6-4) 给出的 MCR 预测的准确度，并令 ACC^{*} 表示由真实的后验概率分布 $P_{Y|X}$ 给出的最大后验概率估计的准确度，则有

$$\mathrm{ACC}^{*}\geqslant\left[1+\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^{2}\right]p_{\min}\geqslant[1+2H_{Y}(\mathbf{f})]p_{\min} \tag{6-12}$$

以及

$$\begin{aligned}\left[\mathrm{ACC}^{*}-\mathrm{ACC}(\mathbf{f})\right]^{2}&\leqslant 2\left[\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^{2}-2H_{Y}(\mathbf{f})\right]p_{\max}\\&\leqslant 2[|\mathcal{Y}|-1-2H_{Y}(\mathbf{f})]p_{\max},\end{aligned} \tag{6-13}$$

其中新定义了 $p_{\min}\triangleq\min_{y\in\mathcal{Y}}P_{Y}(y)$ 及 $p_{\max}\triangleq\max_{y\in\mathcal{Y}}P_{Y}(y)$。

证明 参见附录 D.5。 □

6.3.4　与其他学习问题的联系

本节介绍最大相关回归与信息论及机器学习中若干问题的联系。

6.3.4.1　通用特征选择

首先，最大相关回归提取的特征 $\mathbf{f}$ 是所有 X 的特征中，使得推断 Y 属性的误差最小的属性，即第 3.2 节所介绍的通用特征选择问题的最优特征。

为明确该联系，注意到由 第 3.2 节可知，基于特征 $\mathbf{f}$ 的 n 个独立样本所造成的对 Y 属性的推断误差对应的误差指数正比于单边 H 评分函数 $H_Y(\mathbf{f})$。在此基础上，注意到在基于深度学习框架的最大相关回归中，有

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_{\mathrm{H}} = \arg\max_{\boldsymbol{\theta}} \left[\max_{\mathbf{g}} H(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{g}) \right] = \arg\max_{\boldsymbol{\theta}} H_Y(\mathbf{f}_{\boldsymbol{\theta}}), \tag{6-14}$$

其中前两个等式基于命题 6.2，最后的等式基于单边 H 评分函数与双边 H 评分函数的关系 (参考第 3.5 节中的讨论)。因此，第 6.3.2 节中引入的最大相关回归最优参数 $\boldsymbol{\theta}^*$ 实质上是在所有可能的网络参数 $\boldsymbol{\theta}$ 中使得单边 H 评分函数 $H_Y(\mathbf{f}_{\boldsymbol{\theta}})$ 最大的选择，从而也是使通用特征选择问题中使得错误概率最小化的参数选择。因此，由最大相关回归提取的特征 $\mathbf{f}_{\boldsymbol{\theta}^*}$ 也是参数化函数族 $\mathbf{f}(\cdot;\boldsymbol{\theta})$ 中最小化平均推断误差的最优特征。

6.3.4.2　线性判别分析

从特征空间角度来看，可将最大相关回归所提取的特征解释为最大化类可分度 (Class Separability) 的最优特征，即使得不同标签所对应特征最可分的特征，因此其与 Fisher 提出的线性判别分析 (Linear Discriminant Analysis, LDA)[74] 方法有紧密联系。

具体而言，对给定的数据 X，LDA 求解最优的 X 的线性变换 $\mathbf{f}$ 以最大化类可分度。该可分度通常以平均类内距离或类间距离作为度量，详细的讨论可参考 [75] 的第 10 章。其中，单边 H 评分函数 (6-11) 为较常用的度量之一，其可等价表示为

$$H_Y(\mathbf{f}) = \mathrm{tr}\left\{\mathbf{S}_{\mathrm{t}}^{-1}\mathbf{S}_{\mathrm{b}}\right\} = \mathrm{tr}\left\{\boldsymbol{\Lambda}_{\mathbf{f}}^{-1}\,\mathrm{cov}(\mathbb{E}[\mathbf{f}(X)|Y])\right\}. \tag{6-15}$$

通常称 $\mathbf{S}_{\mathrm{t}} \triangleq \boldsymbol{\Lambda}_{\mathbf{f}}$ 为总散布矩阵 (Total Scatter Matrix) 或混合散布矩阵 (Mixture Scatter Matrix)，$\mathbf{S}_{\mathrm{b}} \triangleq \mathrm{cov}(\mathbb{E}[\mathbf{f}(X)|Y])$ 通常称为类间散布矩阵 (Between-class Scatter Matrix)。$H_Y(\mathbf{f})$ 作为类可分度度量的操作意义由如下命题给出。

命题 6.4： 对给定 X 及 Y，$\mathbf{f}: \mathcal{X} \mapsto \mathbb{R}^k$ 的泛函为单边 H 评分函数 $H_Y(\cdot)$ 当且仅当其满足

(a) 对任意 $\hat{\mathbf{f}}(x) = \mathbf{A}\mathbf{f}(x) + \mathbf{a}$，有 $H_Y(\mathbf{f}) = H_Y(\hat{\mathbf{f}})$，其中 $|\mathbf{A}| \neq 0$ 以及 $\mathbf{A} \in \mathbb{R}^{k\times k}$ $\mathbf{a} \in \mathbb{R}^k$ 为常量，

(b) 若 $\boldsymbol{\Lambda}_{\mathbf{f}} = \mathbf{I}$，则

$$H_Y(\mathbf{f}) = \mathbb{E}\left[\left\|\mathbb{E}\left[\tilde{\mathbf{f}}(X)\middle|Y\right]\right\|^2\right] \tag{6-16a}$$

$$= k - \mathbb{E}\left[\|\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y]\|^2\right]. \tag{6-16b}$$

证明 参见附录 D.6。 □

由于(6-16a) 中的 $\mathbb{E}\left[\left\|\mathbb{E}\left[\tilde{\mathbf{f}}(X)\middle|Y\right]\right\|^2\right]$ 及 $\mathbb{E}\left[\|\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y]\|^2\right]$ 分别对应平均类间距离与平均类内距离，$H_Y(\mathbf{f})$ 可以有效地对类可分度进行度量。

因此，由 (6-14) 可知最大相关回归中最优的 $\mathbf{f}_{\theta^*}$ 是含参模型 $\mathbf{f}_\theta$ 中最大化类可分度的最优特征。与原始 LDA 模型中提取的最佳线性特征相比，基于深度学习的最大相关回归将提取最大类可分度特征的想法推广到非线性特征提取中，具有更为广泛的应用。

6.3.4.3　Softmax 回归

与 第 3.3 节中的讨论对比可知，最大相关回归中使得 $L(\mathbf{f}, \mathbf{g}, b)$ 最小化的最优的 $\mathbf{f}$ 及 $\mathbf{g}$ 分别对应于 Softmax 回归问题中的最优特征及权重，从而对最大相关回归的分析可为神经网络中的特征提取机制的理解提供直接的帮助。

6.4　无监督特征提取

本节考察多模态的无监督学习问题中，从多模态变量中提取具信息性特征的问题，并介绍基于局部信息几何方法的特征提取算法。基于对多变量局部信息几何性质的分析，这里的讨论将 [76] 中对一维最优特征的分析推广到任意 k 维，并且设计了可从高维连续数据样本中提取最优特征的深度学习算法。

6.4.1　信息论意义下的最优无监督特征

给定字母集 $\mathcal{X}^d \triangleq \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ 上的离散随机变量 $X^d = (X_1, \cdots, X_d)$，记其联合分布为 P_{X^d}，则随机变量之间的公共结构可建模为高维隐变量 W，使得 $X_1, \cdots, X_d$ 关于 W 条件独立，亦即 $P_{X^d|W} = \Pi_{i=1}^d P_{X_i|W}$，如图 6.2 所示。以下考察从由 P_{X^d} 生成的独立同分布样本中学习该公共结构的问题。注意到在通常的无监督学习场景中，由于 W 与诸 X_i 之间的相关性通常较为复杂，在标签及诸 X_i 生成模型未知的情形下 W 难以直接辨识结构。基于此，我们转而考虑低维随机变量 U 的学习问题，以使其尽可能包含诸 X_i 间的公共信息，从而可视为学习公共结构 W 具信息性的属性。

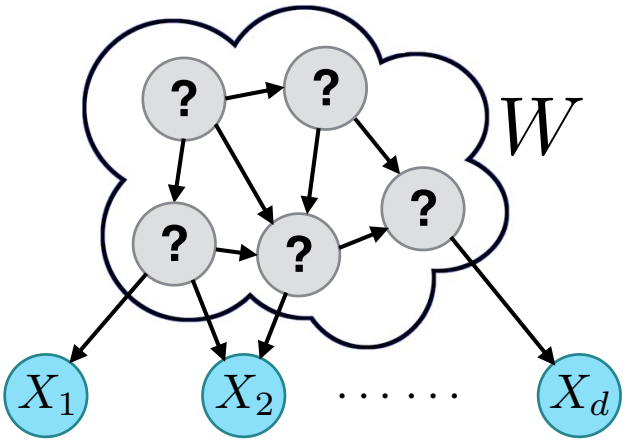


图 6.2　给定包含隐结构信息的变量 W 后随机变量 $X_1, \cdots, X_d$ 条件独立[76]

为刻画这样的 U，将总相关①[77] (Total Correlation) 作为多个随机变量之间的公共信息度量，则属性 U 所包含的有关诸 X_i 公共结构的信息可由给定 U 时总相关减少的程度描述：

$$\mathcal{L}(X^d|U) \triangleq D(P_{X^d}\|P_{X_1}\cdots P_{X_d}) - D(P_{X^d}\|P_{X_1}\cdots P_{X_d}|U).$$

我们希望在满足信息率约束② $I(U;X^d) \leqslant \delta$ 时，求解使得总相关减少量最大的随机变量 U[76]：

$$\underset{P_{UX^d}}{\text{maximize}}\ \mathcal{L}(X^d|U) \tag{6-17a}$$

$$\text{subject to}\ I(U;X^d) \leqslant \delta. \tag{6-17b}$$

特别地，分析中将着重考虑低信息率机制下的 U, 对应于 δ 非常小的情况，从而所求解属性为描述公共结构最具代表性的低维属性。此外，我们自然地假设

$$\min_u P_U(u) > \gamma, \tag{6-18}$$

其中 $\gamma > 0$ 为与 δ 无关的常数。一般而言，优化问题 (6-17) 没有解析解，但在小 δ 机制下，基于局部信息几何方法可将其解表示为某种联合分布矩阵的特征值分解。

① 给定随机变量 $X_1, \cdots, X_d$，总相关定义为联合分布与边缘分布乘积间的 K-L 散度 $D(P_{X_1\cdots X_d}\|P_{X_1}\cdots P_{X_d})$。

② 注意到 $I(U;X^d)$ 度量了 U 所包含关于整个 X^d 的信息总量，而 $\mathcal{L}(X^d|U)$ 度量了公共结构所蕴含的信息。当 δ 取值很小时，引入约束 $I(U;X^d) \leqslant \delta$ 意味着这里重点考察的是 W 的低维属性。

6.4.1.1　局部信息几何分析

为便于陈述，首先基于随机变量的两两联合分布，定义矩阵 $\mathbf{B}$[78]

$$\mathbf{B}=\begin{bmatrix}\mathbf{I}_{(1)} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1d}\\ \mathbf{B}_{21} & \mathbf{I}_{(2)} & \cdots & \mathbf{B}_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{B}_{d1} & \mathbf{B}_{d2} & \cdots & \mathbf{I}_{(d)}\end{bmatrix} \tag{6-19}$$

其中，对所有 i，$\mathbf{I}_{(i)}$ 表示 $|\mathcal{X}_i|\times|\mathcal{X}_i|$ 单位阵；对所有 $i\neq j$，$\mathbf{B}_{ij}$ 为对应的 $(|\mathcal{X}_i|\times|\mathcal{X}_j|)$ 散度转移矩阵 [参见 (2-3)]，其 x_i 行 x_j 列元素为

$$B_{ij}(x_i;x_j)=\frac{P_{X_iX_j}(x_i,x_j)}{\sqrt{P_{X_i}(x_i)}\sqrt{P_{X_j}(x_j)}}.$$

矩阵 $\mathbf{B}$ 的特征值分解有如下性质。

引理 6.1： 令 $\mathbf{B}$ 的特征值与特征向量分别为 $\lambda^{(0)}\geqslant\lambda^{(1)}\geqslant\cdots\geqslant\lambda^{(m-1)}$ 以及 $\boldsymbol{\psi}^{(0)},\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(m-1)}$，其中 $m\triangleq\sum_{i=1}^{d}|\mathcal{X}_i|$ 为 $\mathbf{B}$ 的维度。此外，令 $\mathbf{v}_i$ 为满足 $v_i(x_i)=\sqrt{P_{X_i}(x_i)}$ 的 $|\mathcal{X}_i|$ 维向量，则

1. $\mathbf{B}$ 为半正定矩阵，即 $\lambda^{(m-1)}\geqslant 0$。
2. 其最大特征值 $\lambda^{(0)}=d$，且对应特征向量为 $\boldsymbol{\psi}^{(0)}=\frac{1}{\sqrt{d}}\left[\mathbf{v}_1^{\mathsf{T}},\cdots,\mathbf{v}_d^{\mathsf{T}}\right]^{\mathsf{T}}$。
3. 其第二大特征值 $\lambda^{(1)}\geqslant 1$。
4. 其最小的 $d-1$ 个特征值满足 $\lambda^{(m-d+1)}=\cdots=\lambda^{(m-1)}=0$，且对应的特征空间由形如 $\boldsymbol{\psi}=\left[\alpha_1\mathbf{v}_1^{\mathsf{T}},\cdots,\alpha_d\mathbf{v}_d^{\mathsf{T}}\right]^{\mathsf{T}}$、且诸 α_i 满足 $\sum_{i=1}^{d}\alpha_i=0$ 的向量张成。
5. 对任意 $1\leqslant\ell\leqslant m-d$，若将对应特征向量 $\boldsymbol{\psi}^{(\ell)}$ 分解为诸 $|\mathcal{X}_i|$ 维子向量 $\boldsymbol{\psi}_i^{(\ell)}$，使得

$$\boldsymbol{\psi}^{(\ell)}=\begin{bmatrix}\boldsymbol{\psi}_1^{(\ell)}\\ \vdots\\ \boldsymbol{\psi}_d^{(\ell)}\end{bmatrix}, \tag{6-20}$$

则对任意 i，$\boldsymbol{\psi}_i^{(\ell)}$ 正交于 $\mathbf{v}_i$。

证明 参见附录 D.7。 □

为便于分析，定义矩阵 $\tilde{\mathbf{B}}$ 为

$$\tilde{\mathbf{B}}\triangleq\mathbf{B}-d\cdot\boldsymbol{\psi}^{(0)}(\boldsymbol{\psi}^{(0)})^{\mathsf{T}}. \tag{6-21}$$

则由引理 6.1 知，$\tilde{\mathbf{B}}$ 特征值为 $\lambda^{(1)} \geqslant \cdots \geqslant \lambda^{(m-d)} \geqslant 0 = \lambda^{(m-d+1)} = \cdots = \lambda^{(m)}$，对应特征向量 $\boldsymbol{\psi}^{(1)}, \cdots, \boldsymbol{\psi}^{(m-1)}, \boldsymbol{\psi}^{(0)}$。此外，定义一系列函数 $f_i^{(\ell)} \colon \mathcal{X}_i \mapsto \mathbb{R}$ 为

$$f_i^{(\ell)}(x_i) = \frac{\psi_i^{(\ell)}(x_i)}{\sqrt{P_{X_i}(x_i)}}, \qquad \forall\, i, \ell, \tag{6-22}$$

其中 $\boldsymbol{\psi}_i^{(\ell)}$ 为由 (6-20) 给出的 $\boldsymbol{\psi}^{(\ell)}$ 的第 i 个子向量。于是，结合 引理 6.1 及 (6-22) 可知 $f_i^{(\ell)}(X_i)$ 为零均值函数且 $\sum_{i=1}^{d} \mathbb{E}\left[(f_i^{(\ell)}(X_i))^2\right] = 1$。此外，这些函数可导出关于 U, X^d 联合分布的一个指数族，定义如下。

定义 6.3： 令 $\mathcal{H}$ 为满足零均值及单位方差的函数 $h \colon \mathcal{U} \mapsto \mathbb{R}$ 的集合，则关于 U, X^d 的指数族 $\mathcal{P}_{\exp}^{(\delta)}$ 定义为

$$\mathcal{P}_{\exp}^{(\delta)} = \left\{ \frac{1}{Z} P_U(u) P_{X^d}(x^d) \cdot \exp\left(\sqrt{2\delta} \frac{h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right) \colon h \in \mathcal{H} \right\},$$

其中 Z 为归一化因子。

对给定 X^d，该指数族同时定义了由 X^d 生成的、联合分布取自 $\mathcal{P}_{\exp}^{(\delta)}$ 的一族随机变量 U。以下定理表明，在小 δ 机制下优化问题 (6-17) 的最优解可由该指数族刻画。

定理 6.2： 优化问题 (6-17a) 最优值为

$$\max_{P_{UX^d}} \mathcal{L}(X^d | U) = \delta \left(\lambda^{(1)} - 1 \right) + o(\delta), \tag{6-23}$$

且最优值可由 $\mathcal{P}_{\exp}^{(\delta)}$ 中的分布取得。此外，对任意使 (6-23) 最优的 P_{UX^d}，存在分布 $\hat{P}_{UX^d} \in \mathcal{P}_{\exp}^{(\delta)}$ 使得对任意 $(u, x^d) \in \mathcal{U} \times \mathcal{X}^d$ 都有

$$\left| P_{UX^d}(u, x^d) - \hat{P}_{UX^d}(u, x^d) \right| = o\left(\sqrt{\delta} \right).$$

证明 参见附录 D.8。 □

根据定理 6.2，由 X^d 及 $\mathcal{P}_{\exp}^{(\delta)}$ 给出的诸随机变量 U 为包含 X^d 间公共结构信息量最大的诸属性。该信息可由基于 X^d 估计 U 的对数似然函数的导出：

$$\log \frac{P_{X^d|U=u}(x^d)}{P_{X^d}(x^d)} = \frac{\sqrt{2\delta} h(u)}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) + o\left(\sqrt{\delta} \right). \tag{6-24}$$

由上式可知，对数似然函数在不同的 $U = u$ 取值时幅度可能不同 (由于 $h(u)$ 不同)，但都正比于函数 $\sum_{i=1}^{d} f_i^{(1)}(x_i)$，从而函数空间 $\left\{ \alpha \sum_{i=1}^{d} f_i^{(1)}(x_i) \colon \alpha \in \mathbb{R} \right\}$ 可解

释为 X^d 的函数空间中关于其公共结构最具信息性的一维子空间。注意该最优子空间的求解过程与线性主成分分析[79]类似，区别在于这里考虑的函数空间更具一般性。之后进一步的分析可表明，这里求解最优子空间的方法可视为经典主成分分析的一种非线性推广。

此外，由于 $\boldsymbol{\psi}^{(1)}$ 为 $\mathbf{B}$ 的第二个特征向量，故其在所有与 $\boldsymbol{\psi}^{(0)}$ 正交的单位向量中使得 $\boldsymbol{\psi}^{\mathsf{T}}\mathbf{B}\boldsymbol{\psi}$ 最大化。因此，由 (6-22) 所定义的 $f_i^{(1)}(X_i)$ 为如下联合相关优化问题的最优解：

$$\begin{aligned} \underset{f_i:\ \mathcal{X}_i\mapsto\mathbb{R},\ i=1,\cdots,d}{\text{maximize}} \quad & \mathbb{E}\left[\sum_{i\neq j} f_i(X_i)f_j(X_j)\right] \\ \text{subject to} \quad & \mathbb{E}\left[f_i(X_i)\right]=0, \quad i=1,\cdots,d \\ & \mathbb{E}\left[\sum_{i=1}^{d} f_i^2(X_i)\right]=1, \quad i=1,\cdots,d. \end{aligned}$$

故求解函数表示 $f_i^{(1)}(X_i), i=1,\cdots,d$ 本质上是在各 X_i 对应的一维函数子空间中，使得这些子空间联合相关最大的选择。因此，这些子空间及对应的函数表示包含较多关于随机变量间公共结构的信息。

6.4.1.2　具信息性的 k 维属性

除第一个特征向量 $\boldsymbol{\psi}^{(1)}$ 外，$\tilde{\mathbf{B}}$ 的其它特征向量实质上给出了描述公共结构的最优 k 维属性的函数表示。为此，考虑关于 k 维属性 $U^k=(U_1,\cdots,U_k)$ 的如下优化问题①：

$$\underset{P_{U^kX^d}}{\text{maximize}} \quad \mathcal{L}\left(X^d|U^k\right), \tag{6-25}$$

其中 $\mathcal{L}\left(X^d|U^k\right)$ 定义由 (6.4.1) 给出，且作为优化变量的联合分布 $P_{U^kX^d}$ 满足如下约束：1) 对任意 $i=1,\cdots,k$，U_i 取自集合 $\mathcal{U}_i$；2) $\delta\geqslant I(U_1;X^d)\geqslant\cdots\geqslant I(U_k;X^d)$；3) 对任意 $i=1,\cdots,d$，存在某个与 δ 无关的常数 $\gamma>0$，使得 $\min_{u_i\in\mathcal{U}_i}P_{U_i}(u_i)>\gamma$ 成立；4) $U_1,\cdots,U_k$ 相互独立；5) $U_1,\cdots,U_k$ 关于 X^d 条件独立。

为求解优化问题(6-25), 首先引入与 k 维属性相关的指数族如下。

① 作为对比，CorEx[29] 对给定 U_i 字母集大小，通过求解如下优化问题 [参见[29] 中的等式 (4)] 选择属性 $(U_1,\cdots,U_k)$：

$$\underset{X_{G_i},P_{U_i|X_{G_i}},\ i=1,\cdots,k}{\text{maximize}} \quad \sum_{i=1}^{k}\mathcal{L}\left(X_{G_i}|U_i\right),$$

其中 $\{X_{G_1},\cdots,X_{G_k}\}$ 为集合 $\{X_1,\cdots,X_d\}$ 的分划。

定义 6.4： 给定 $i=1,\cdots,k$，令 $\mathcal{H}_i$ 为满足零均值及单位方差的函数所构成的集合，则关于 U^k, X^d 的指数族 $\mathcal{P}^{(\delta)}_{\exp,k}$ 定会为

$$\mathcal{P}^{(\delta)}_{\exp,k}=\left\{\frac{1}{Z_k}\left[\prod_{j=1}^{k}P_{U_i}(u_i)\right]P_{X^d}(x^d)\cdot\exp\left(\sqrt{2\delta}\sum_{\ell=1}^{k_0}h_\ell(u_\ell)\sum_{j=1}^{k_0}\frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}}\sum_{i=1}^{d}f_i^{(j)}(x_i)\right)\right.$$
$$\left.:\ h_\ell\in\mathcal{H}_\ell,\mathbf{Q}=[q_{ij}]_{k_0\times k_0},\mathbf{Q}^{\mathsf{T}}\mathbf{Q}=\mathbf{I}_{k_0}\right\},$$

其中 $f_i^{(j)}(x_i)$ 定义由 (6-22) 给出，Z_k 为归一化因子，$k_0=\min\{k,k^*\}$ 且

$$k^*\triangleq\max\left\{i:\ \lambda^{(i)}>1\right\}.$$

从而，优化问题(6-25)的最优解可用指数族 $\mathcal{P}^{(\delta)}_{\exp,k}$ 表述如下。

定理 6.3： 优化问题 (6-25) 的最优值为

$$\max_{P_{U^kX^d}}\mathcal{L}\left(X^d|U^k\right)=\delta\left(\sum_{\ell=1}^{k_0}\lambda^{(\ell)}-k_0\right)+o(\delta),\tag{6-26}$$

且最优解可由 $\mathcal{P}^{(\delta)}_{\exp,k}$ 中分布取得。此外，对任意取得最优值 (6-26)的分布 $P_{U^kX^d}$，存在分布 $\hat{P}_{U^kX^d}\in\mathcal{P}^{(\delta)}_{\exp,k}$ 使得对任意 $(u^k,x^d)\in\mathcal{U}_1\times\cdots\times\mathcal{U}_k\times\mathcal{X}^d$，有

$$\left|P_{U^kX^d}(u^k,x^d)-\hat{P}_{U^kX^d}(u^k,x^d)\right|=o\left(\sqrt{\delta}\right).$$

证明 参见附录 D.9。 □

由定义 6.4 及定理 6.3 可知，当 $k>k^*$ 时，最优联合分布 $P_{U^{k^*}X^d}$ 须取自 $\mathcal{P}^{(\delta)}_{\exp,k^*}$，并令最后 $k-k^*$ 属性 $U_{k^*+1},\cdots,U_k$ 与 X^d 独立。因此，仅有前 k^* 个属性可有效降低总相关，该结论本质上建立了实际问题中属性维度 k 的确定准则。

注意到由 定义 6.4 可知最优的属性所对应的对数似然函数为 $\sum_{i=1}^{d}f_i^{(\ell)}(x_i),\ell=1,\cdots,k$，该结果可视为对 (6-24) 的 k 维推广。此外，根据附录 D.10 可知由 (6-22) 定义的 $f_i^{(\ell)}$ 为如下问题的最优解：

$$\underset{\underline{f}_i:\ \mathcal{X}_i\mapsto\mathbb{R}^k,\ i=1,\cdots,d}{\text{maximize}}\quad\mathbb{E}\left[\sum_{i\neq j}\underline{f}_i^{\mathsf{T}}(X_i)\underline{f}_j(X_j)\right]\tag{6-27a}$$

$$\text{subject to}\quad\mathbb{E}\left[\underline{f}_i(X_i)\right]=\underline{0},\ \forall\, i\tag{6-27b}$$

$$\mathbb{E}\left[\sum_{i=1}^{d}\underline{f}_i(X_i)\underline{f}_i^{\mathsf{T}}(X_i)\right]=\mathbf{I}_k,\tag{6-27c}$$

其中 $\mathbf{I}_k$ 为 k 维单位阵。故对 $i=1,\cdots,d$，函数表示 $f_i^{(\ell)}(X_i), \ell=1,\cdots,k$ 建立了 X_i 的 k 维函数子空间，以使得不同的 X_i 所对应子空间之间的联合相关最大化。

例 6.1 (二进制序列中的模式提取)： 令 $b_1,\cdots,b_r \in \{1,-1\}$ 为彼此独立的 $\mathsf{Bern}(\frac{1}{2})$ 比特位，构造随机变量 $X_i = b_{\mathcal{I}_i} \triangleq (b_j)_{j\in\mathcal{I}_i}$ 对应该随机比特的子集，其中 $\mathcal{I}_i \subseteq \{1,\cdots,r\}$ 为对应的指标集，则前述信息论方法实质上将提取出随机变量 X^d 中出现频率最高的比特模式。为方便陈述结果，定义 $w(\mathcal{I})$ 为集合 $\mathcal{I}_i$ $(i=1,\cdots,d)$ 中包含 $\mathcal{I}$ 的集合的数量，即

$$w(\mathcal{I}) \triangleq \sum_{i=1}^{d} \mathbb{1}_{\{\mathcal{I}\subset\mathcal{I}_i\}}. \tag{6-28}$$

此外，记 $\varnothing = \mathcal{J}_0,\cdots,\mathcal{J}_{2^r-1}$ 为 $\{1,\cdots,r\}$ 的 2^r 个子集，使其满足 $d = w(\mathcal{J}_0) \geqslant w(\mathcal{J}_1) \geqslant \cdots \geqslant w(\mathcal{J}_{2^r-1})$。于是由附录 D.11 可知，该问题所对应矩阵 $\mathbf{B}$ 的特征值为

$$\lambda^{(\ell)} = w(\mathcal{J}_\ell), \quad \ell = 0,\cdots,m-1, \tag{6-29}$$

其中 $m = \sum_{i=1}^{d} 2^{|\mathcal{I}_i|}$ 为矩阵 $\mathbf{B}$ 的维度。因此，矩阵 $\mathbf{B}$ 的特征值 $\lambda^{(\ell)}$ 实质上等于相应的比特模式 $b_{\mathcal{J}_\ell}$ 出现在诸 X_i 中的次数，故最大特征值对应出现频率最高的比特模式。此外对于 $\lambda^{(\ell)}>0$，相应的由 (6-22) 所定义的函数 $f_i^{(\ell)}(X_i)$ $(i=1,\cdots,d)$ 为

$$f_i^{(\ell)}(X_i) = \begin{cases} \dfrac{1}{\sqrt{w(\mathcal{J}_\ell)}} \displaystyle\prod_{j\in\mathcal{J}_\ell} b_j & \text{若} \mathcal{J}_\ell \subset \mathcal{I}_i \\ 0 & \text{其他情况.} \end{cases} \tag{6-30}$$

因此，X^d 的第 ℓ 个最优函数表示 (参见第 6.4.1.2 节) 为

$$\sum_{i=1}^{d} f_i^{(\ell)}(X_i) = \sqrt{w(\mathcal{J}_\ell)} \prod_{j\in\mathcal{J}_\ell} b_j,$$

其仅取决于 $\mathcal{J}_\ell$ 中的那些比特位。

为具体阐释前述性质，构造 $r=d=3$，$X_1=\{b_1,b_2\}, X_2=\{b_2,b_3\}$，且 $X_3=\{b_1,b_3\}$，则对 $\{1,2,3\}$ 的所有子集，(6-28) 所定义的函数 $w(\cdot)$ 的取值分别为

$$w(\varnothing)=3, \quad w(\{1\})=w(\{2\})=w(\{3\})=2,$$
$$w(\{1,2\})=w(\{2,3\})=w(\{3,1\})=1, \quad w(\{1,2,3\})=0.$$

故相应的 $\mathbf{B}$ 的该特征值为

$$\lambda^{(0)}=3, \quad \lambda^{(1)}=\lambda^{(2)}=\lambda^{(3)}=2,$$

$$\lambda^{(4)} = \lambda^{(5)} = \lambda^{(6)} = 1, \quad \lambda^{(7)} = 0.$$

此外，对应的 $f_i^{(\ell)}(X_i)$ 将满足

$$\sum_{i=1}^{3} f_i^{(\ell)}(X_i) = \sqrt{2} b_\ell, \quad \ell = 1, 2, 3,$$

以及

$$\sum_{i=1}^{3} f_i^{(\ell)}(X_i) = \begin{cases} b_1 b_2 & \ell = 4 \\ b_2 b_3 & \ell = 5 \\ b_3 b_1 & \ell = 6. \end{cases}$$

6.4.2　最优特征提取算法

基于前述具信息性函数表示的信息论分析，本节介绍可用于实践的最优函数算法。给定 $X_1, \cdots, X_d$ 的数据样本，一个自然的想法为：计算样本经验分布所对应的 **B** 矩阵，并求其特征值分解。但该方法难以实践因为：(1) 样本数不足以准确地估计相应的联合分布；(2) 实际数据所对应 **B** 矩阵维度过大，无法直接计算特征值分解。

6.4.2.1　多元交替条件期望算法

除直接计算特征值分解的方法外，可使用幂法[80] 完成矩阵的特征向量的高效求解。对给定矩阵及初始向量，幂法重复执行矩阵向量间的乘法操作，且若所有特征值非负，幂法可线性收敛至最大特征值所对应的特征向量。为使用幂法计算 **B** 的第二个特征向量，这里初始向量选为 $\boldsymbol{\psi} = \begin{bmatrix} \boldsymbol{\psi}_1^\mathsf{T} & \cdots & \boldsymbol{\psi}_d^\mathsf{T} \end{bmatrix}^\mathsf{T}$ 并使得对任意 i，$\boldsymbol{\psi}_i$ 与 $\mathbf{v}_i$ 正交。该条件可保证 $\boldsymbol{\psi}$ 正交于 $\boldsymbol{\psi}^{(0)}$。又因 **B** 半正定，在 $\boldsymbol{\psi}$ 与第二个特征向量 $\boldsymbol{\psi}^{(1)}$ 不正交的条件下，$\boldsymbol{\psi}$ 将收敛至 $\boldsymbol{\psi}^{(1)}$。具体地，重复执行矩阵乘法 $\boldsymbol{\psi} \leftarrow \mathbf{B}\boldsymbol{\psi}$，或等价地表示为：对所有 i，

$$\boldsymbol{\psi}_i \leftarrow \boldsymbol{\psi}_i + \sum_{j \neq i} \mathbf{B}_{ij} \boldsymbol{\psi}_j. \tag{6-31}$$

若将 $\boldsymbol{\psi}_i$ 视为向量并引入对应函数 $f_i(x_i) = \psi_i(x_i)/\sqrt{P_{X_i}(x_i)}$，则根据命题 2.3 可将 (6-31) 等价表示为对函数的条件期望操作：

$$f_i(X_i) \leftarrow f_i(X_i) + \mathbb{E}\left[\sum_{j \neq i} f_j(X_j) \middle| X_i \right], \tag{6-32}$$

算法 7 多元交替条件期望 (Multivariate ACE，MACE) 算法[76]

Require: 随机变量 $X_1, \cdots, X_d$ 的数据样本 $\underline{x}^{(\ell)} = (x_1^{(\ell)}, \cdots, x_d^{(\ell)}), \ell = 1, \cdots, n$

1: *初始化*：随机选取零均值函数 $\vec{f} = (f_1, \cdots, f_d)$.

2: **repeat**

3: 　求解交替条件期望：$f_i(X_i) \leftarrow f_i(X_i) + \mathbb{E}\left[\sum_{j\neq i} f_j(X_j)\middle| X_i\right]$.

4: 　归一化：$f_i(X_i) \leftarrow f_i(X_i) / \sqrt{\mathbb{E}\left[\sum_{i=1}^{d} f_i^2(X_i)\right]}$.

5: **until** $\mathbb{E}\left[\sum_{i\neq j} f_i(X_i) f_j(X_j)\right]$ 值停止增长。

算法 8 $\vec{f}^{(k)}$ 的计算[76]

Require: 随机变量 $X_1, \cdots, X_d$ 的样本 $\underline{x}^{(i)} = (x_1^{(i)}, \cdots, x_d^{(i)}), i = 1, \cdots, n$，以及计算所得函数 $\vec{f}^{(1)}, \cdots, \vec{f}^{(k-1)}$。

1: *初始化*：随机选取零均值函数 $\vec{f}^{(k)} = (f_1^{(k)}, \cdots, f_d^{(k)})$.

2: **repeat**

3: 　对 $\vec{f}^{(k)}$ 执行算法 7 的第 3 至 4 行的操作.

4: 　Gram-Schmidt 正交化过程：$\vec{f}^{(k)} \leftarrow \vec{f}^{(k)} - \sum_{\ell=1}^{k-1} \langle \vec{f}^{(\ell)}, \vec{f}^{(k)} \rangle \cdot \vec{f}^{(\ell)}$

5: **until** $\mathbb{E}\left[\sum_{i\neq j} f_i^{(k)}(X_i) f_j^{(k)}(X_j)\right]$ 值停止增长。

故幂法可转化为相应的交替条件期望算法[42]，如算法 7 所示。可通过算法计算第 6.4.1 节中导出的最优函数表示，其中 $\boldsymbol{\psi}_i$ 与 $\mathbf{v}_i$ 正交约束对应于算法初始化步骤中对函数零均值的要求。

6.4.2.2　基于特征值分解的 k 维函数表示求解

可进一步推广算法 7 将其用于前 k 个特征向量 $\boldsymbol{\psi}^{(1)}, \cdots, \boldsymbol{\psi}^{(k)}$ 及相应函数表示的计算。为描述该计算过程，首先将第 ℓ 个函数表示记为 $\vec{f}^{(\ell)} = (f_1^{(\ell)}, \cdots, f_d^{(\ell)})$，其中 $f_i^{(\ell)}$ 定义由 (6-22) 给出。则因对 $\ell \leqslant k-1$，$\boldsymbol{\psi}^{(k)}$ 都正交于 $\boldsymbol{\psi}^{(\ell)}$，可类似于 $\vec{f}^{(1)}$ 使用幂法计算第 k 个函数表示 $\vec{f}^{(k)}$，但需引入额外的正交性约束

$$\left\langle \vec{f}^{(\ell)}, \vec{f}^{(k)} \right\rangle \triangleq \sum_{i=1}^{d} \mathbb{E}\left[f_i^{(\ell)}(X_i) f_i^{(k)}(X_i)\right] = 0, \quad \ell \leqslant k-1$$

以保证其与前 $k-1$ 个函数表示的正交性。故可结合算法 7 中的幂法及 Gram–Schmidt 正交化过程用于计算 $\vec{f}^{(k)}$，具体过程如算法 8 所示。注意算法 7 及 算法 8 的计算复杂度均与数据集大小呈线性关系，从而比直接计算 **B** 奇异值分解要更为高效。

6.4.2.3　高维数据中的最优函数表示提取

尽管一般情况下算法 8 所需样本数比直接估计联合分布以及矩阵 $\mathbf{B}$ (或 $\tilde{\mathbf{B}}$) 要少，为得到相对准确的条件期望估计 (6-32)，训练样本数仍需要达到 X_i 字母集大小的量级。该样本数的要求在实际处理高维或连续数据时往往难以满足。针对该应用场景，下面建立基于深度神经网络的具信息性函数表示的学习框架，其中关键思路为：根据 Eckart–Young–Mirsky 定理[56]，$\tilde{\mathbf{B}}$ 的前 $k \leqslant m$ 个特征向量可通过以下低秩恢复问题计算：

$$\mathbf{\Psi}^* = \underset{\mathbf{\Psi}\in\mathbb{R}^{m\times k}}{\arg\min} \left\|\tilde{\mathbf{B}} - \mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\right\|_{\mathrm{F}}^2 \tag{6-33}$$

其中 $\mathbf{\Psi}^*$ 的列空间为 $\tilde{\mathbf{B}}$ 前 k 特征向量张成的子空间. 由无约束优化问题 (6-33) 可给出基于神经网络的具信息性函数提取所对应的损失函数。

命题 6.5： 对任意 $i = 1, \cdots, d$，令 $\mathbf{\Psi}_i$ 表示相应的 $|\mathcal{X}_i| \times k$ 子矩阵使得 $\mathbf{\Psi} = \begin{bmatrix}\mathbf{\Psi}_1^{\mathsf{T}} & \cdots & \mathbf{\Psi}_d^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}$，并定义 k 维函数 $\underline{f}_i : \mathcal{X}_i \mapsto \mathbb{R}^k$ 为 $\underline{f}_i(x_i) = \mathbf{\Psi}_i^{\mathsf{T}}(x_i)/\sqrt{P_{X_i}(x_i)}$，其中 $\mathbf{\Psi}_i(x_i)$ 表示矩阵 $\mathbf{\Psi}_i$ 的第 x_i 行。则

$$\left\|\tilde{\mathbf{B}} - \mathbf{\Psi}\mathbf{\Psi}^{\mathsf{T}}\right\|_{\mathrm{F}}^2 = \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - 2H\left(\underline{f}_1(X_1), \cdots, \underline{f}_d(X_d)\right), \tag{6-34}$$

其中

$$H\left(\underline{f}_1(X_1), \cdots, \underline{f}_d(X_d)\right) \triangleq \sum_{i=1}^{d}\sum_{j=1}^{d} H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right),$$

且对任意 i, j，$H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right)$ 定义为

$$\begin{aligned} H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right) \triangleq\ & \mathbb{E}\left[\underline{f}_i^{\mathsf{T}}(X_i)\underline{f}_j(X_j)\right] - \left(\mathbb{E}\left[\underline{f}_i(X_i)\right]\right)^{\mathsf{T}}\mathbb{E}\left[\underline{f}_j(X_j)\right] \\ & - \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathsf{T}}(X_i)\right]\mathbb{E}\left[\underline{f}_j(X_j)\underline{f}_j^{\mathsf{T}}(X_j)\right]\right\}. \end{aligned}$$

证明 参见附录 D.12. □

注意到对零均值的 $\underline{f}_i(X_i)$ 及 $\underline{f}_j(X_j)$，$H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right)$ 定义与定义 3.4 中的 H 评分函数一致，因此这里将 $H\left(\underline{f}_1(X_1), \cdots, \underline{f}_d(X_d)\right)$ 称为*多元 H 评分函数* (Multivariate H-score, MH-score)，则由 (6-34) 求解具信息性函数表示的优化问题 (6-33) 等价于

$$\max_{f_i:\ \mathcal{X}_i\mapsto\mathbb{R}^k,\ i=1,\cdots,d} H\left(\underline{f}_1(X_1), \cdots, \underline{f}_d(X_d)\right), \tag{6-35}$$

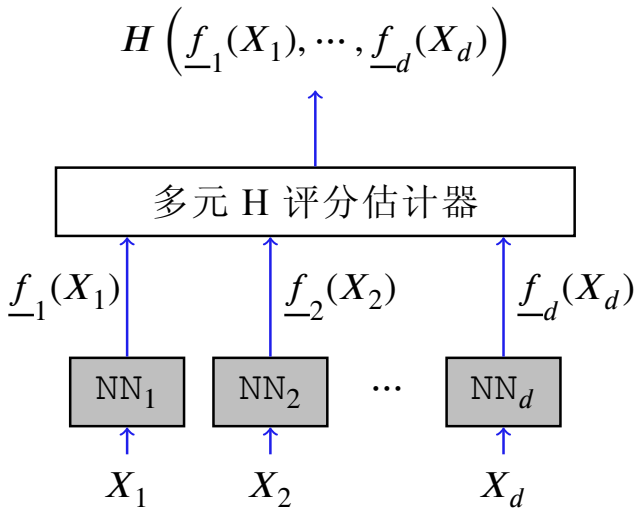


图 6.3　估计最优函数表示的神经网络结构，其中第 i 个子网络 NN_i 用于从输入 X_i 中提取特征 $\underline{f}_i(X_i)$。

此外，由于 H 评分函数可从数据样本中高效地求得 (具体算法可参考 [71] 中的算法 1), 多元 H 评分函数亦可有效地从数据样本中估计得出。基于此，可由优化问题 (6-35) 得出神经网络训练框架如下：给定 $X_1,\cdots,X_d$ 的训练样本，设计 d 个神经网络，其中神经网络 NN_i 以 X_i 为输入，生成表示 $\underline{f}_i(X_i)$。接着，训练神经网络权重以最小化负的多元 H 评分函数，则具信息性的函数表示由训练所得的 d 个神经网络生成，如图 6.3 所示。与 MACE 算法相比，该方法借助了深度神经网络在特征表示方面的优势，从而可针对输入数据的特性进行高效的最优特征提取。

6.4.3　与其他机器学习问题的联系

本节给出所求最优函数表示与 HGR 最大相关问题，线性主成分分析[79] 及一致函数映射[81] 的联系，从而可为已有机器学习算法提供信息论角度的解释。

6.4.3.1　HGR 最大相关问题

为说明这里无监督特征提取方法与 HGR 最大相关问题的联系，注意若 $d=2$ 则 第 6.4.1 节给出的最优函数恰好是两个随机变量间的最大相关函数 (可参考定义 2.4)。此外，一般情况下求解最优函数表示的问题可视为对最大相关的多元推广 [对比 (6-27) 的优化问题]：

定义 6.5： 对取值在离散集合 $\mathcal{X}_i$ ($i=1,\cdots,d$) 上的联合随机变量 $X_1,\cdots,X_d$，广义最大相关定义为

$$\rho^*(X_1,\cdots,X_d)\triangleq\max\frac{1}{d-1}\mathbb{E}\left[\sum_{i\neq j}f_i(X_i)f_j(X_j)\right] \tag{6-36}$$

其中对所有 i，函数 $f_i: \mathcal{X}_i \mapsto \mathbb{R}$ 满足约束 $\mathbb{E}\left[f_i(X_i)\right]=0$, $\mathbb{E}\left[\sum_{i=1}^{d} f_i^2(X_i)\right]=1$。易验证相关系数满足 $0 \leqslant \rho^*(X_1,\cdots,X_d) \leqslant 1$，且 $\rho^*(X_1,\cdots,X_d)=0$ 当且仅当随机变量 $X_1,\cdots,X_d$ 两两独立。

除上述定义外，之前的若干工作也研究了最大相关性的多元推广。如网络最大相关[82](Network Maximal Correlation, NMC) 定义了类似于 (6-36) 的相关性度量，不同之处在于[82] 的约束条件为 $\mathbb{E}\left[f_i(X_i)\right]=0$, $\mathbb{E}\left[f_i^2(X_i)\right]=1$, $\forall i$。此外，[83] 定义了最大相关主成分分析 (Maximally Correlated Principal Component Analysis, MCPCA) 方法，并说明了在特定情况下 MCPCA 的解也对应 $\tilde{\mathbf{B}}$ 的第一个特征向量 (参见 [83] 中的定理 5)。与这些工作相比，这里提出的方法本质上通过研究提取多随机变量共同结构的问题，给出了信息论意义上对最大相关性的推广，且可指导实用算法设计。

6.4.3.2　线性主成分分析

第 6.4.1 节中所研究的函数表示也可看作是线性主成分分析[79] 的非线性推广。为说明该联系，考察多维数据样本 $\underline{x}^{(\ell)}=(x_1^{(\ell)},\cdots,x_d^{(\ell)}) \in \mathbb{R}^d$, $\ell=1,\cdots,n$。不失一般性，设该多维样本每个维度都满足零均值及单位方差的条件，亦即对所有 i，$\sum_{\ell=1}^{n} x_i^{(\ell)}=0$, $\frac{1}{n}\sum_{\ell=1}^{n}(x_i^{(\ell)})^2=1$。则线性主成分分析通过求解单位向量 $\underline{w}=(w_1,\cdots,w_d)$ 以最大化 $\sum_{\ell=1}^{n}\langle \underline{w}, \underline{x}^{(\ell)}\rangle^2$，该问题等价于在约束条件

$$1=\sum_{i=1}^{d} w_i^2=\sum_{i=1}^{d}\mathbb{E}\left[\left(w_i X_i\right)^2\right], \tag{6-37}$$

下最大化函数

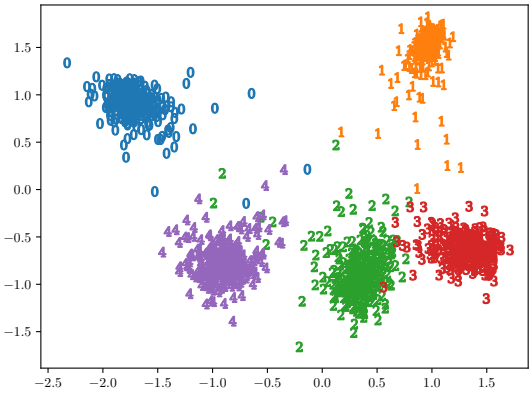
$$\frac{1}{n}\sum_{\ell=1}^{n}\sum_{i\neq j}\left(w_i x_i^{(\ell)}\right)\left(w_j x_j^{(\ell)}\right)=\mathbb{E}\left[\sum_{i\neq j}\left(w_i X_i\right)\cdot\left(w_j X_j\right)\right], \tag{6-38}$$

其中 (6-38) 与 (6-37) 的期望分别取在样本经验分布 $P_{X_iX_j}$ 及 P_{X_i} 上。对照定义 6.5可知，这里研究的最优函数表示将线性主成分分析推广到了一般的函数空间。需要指出的是，[33] 通过局部几何方法，对服从 Gaussian 分布的数据给出了主成分分析的非线性推广，而此处研究的方法本质上为主成分分析在一般离散数据上的另一个推广。

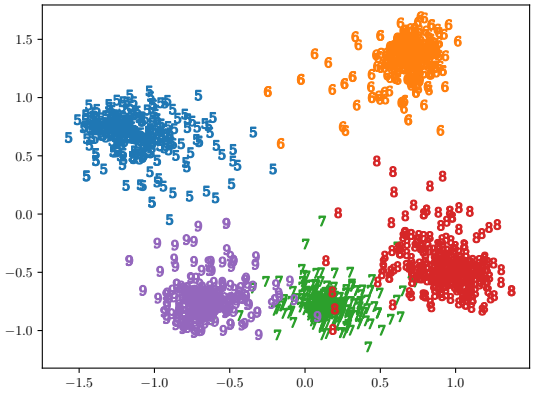
6.4.3.3　一致函数映射

给定一系列三维形状，从中提取对形状描述的公共成分是计算机视觉中的典型问题。例如，给定不同形态的三维椅子模型以及形态间的 (有噪) 映射，从中恢

(a) 数据样本



(b) 估计所得 $\mathbf{f}(X)$



(c) 估计所得 $\mathbf{g}(Y)$

图 6.4　用于实验的 MNIST 图像对及所导出的二维最大相关函数

复关于椅子的结构信息。一致函数映射 (Consistent Functional Map)[81,84-85] 是提取该公共结构的一个有效方法，其主要思路为将公共结构建模为三维形状所在函数空间的低维子空间，并将给定的形状间的有噪映射建模为所对应函数空间之间的转移映射 (Transition Map)。在此基础上，一致函数映射方法希望求解不同形状函数空间的低维子空间，且该子空间为任意到自身的转移映射的不变子空间。

一致函数映射与前述特征提取方法类似，且形状间转移映射类比于随机变量间的映射。实际上，若将形状 i 与形状 j 间有噪映射记为 $\mathbf{M}_{ij}$ [参见 [81] 式 (8) 中的 $X_{ij}^{\star}$]，只需将 (6-19) 中 $\mathbf{B}$ 的块 $\mathbf{B}_{ij}$ 替换为形状间的有噪映射 $\mathbf{M}_{ij}$，则一致函数映射方法所求子空间可表示对应矩阵的特征值分解。因此，前述特征提取方法可视为对一致函数映射方法在随机变量上的拓展，其应用范围将更广泛。

6.5　实验结果

为验证所设计算法的性能，本节在实际数据集上开展了一系列相关实验。

6.5.1　最大相关函数提取

本实验中考察从统计相关的图像中计算最大相关函数。为此，令 X 和 Y 均为 MNIST 数据集的图像，且对应标签满足 $l(X) \in \{0,1,2,3,4\}$ 及 $l(Y) = l(X) + 5$，如图 6.4(a) 所示。由于 X 与 Y 对应标签满足一一映射关系，故 X 与 Y 的统计关联性很强。具体地，为生成这样的样本对，首先将数据集所有中所有标签为 $\{0,1,2,3,4\}$

表 6.1　由优化 H 评分函数及 CA-NN[87] 分别计算所得最大相关函数的分类准确率 [%]

有效维度①	2	3	4
优化 H 评分函数所得 **f**	**98.7**	**99.3**	**99.8**
CA-NN 所得 **f**	97.0	98.4	99.7
优化 H 评分函数所得 **g**	**97.5**	**98.7**	**99.4**
CA-NN 所得 **g**	94.8	97.5	99.2

① 有效维度定义为所提取的相关模式的数量，等同于优化 H 评分函数方法中 $\mathrm{NN}_{\mathbf{f}}$ 与 $\mathrm{NN}_{\mathbf{g}}$ 输出特征的维度。但由于 CA-NN 所提取的特征包含常数函数所对应的平凡模式，因此其有效维度比特征维度要少 1 维，即：为取得有效维度 k，CA-NN 网络输出维度应设为 $k+1$。

的样本作为 $\{x_i\}_{i=1}^n$；在此基础上，对每个 x_i，在数据集中随机选择标签为 $[l(x_i)+5]$ 的样本作为 y_i；最后将这些 (x_i, y_i) 随机分为数据集与训练集。

接着采用图 6.5 所示卷积神经网络作为 $\mathrm{NN}_{\mathbf{f}}$ 与 $\mathrm{NN}_{\mathbf{g}}$ 以提取 $k=2$ 维最大相关函数，使用 ADADELTA[86] 为优化器训练整个网络 100 期，其中小批量大小设为 128。所提取的最大相关函数 **f** 及 **g** 分别如图 6.4(b) 及图 6.4(c) 所示。从图中可以发现，尽管训练过程不依赖于标签，所提取特征自然包含了标签 $l(x_i)$ 与 $l(y_i)$ 的信息。为进一步量化提取特征的性能，将特征 $\mathbf{f}(x_i)$ 送入单层神经网络构成的线性的分类器 (等价于 Softmax 回归) 用于预测标签 $l(x_i)$，可得测试准确率为 98.7%。与之类似，基于 $\mathbf{g}(y_i)$ 预测 $l(y_i)$ 的测试准确率为 97.5%。

进一步展开 $k=3,4$ 的实验，并将结果与同样用于求解最大相关函数的方法 CA-NN (Correspondence Analysis Neural Network)[87] 作对比，相应结果由表 6.1 给出。由表中结果可知，与 CA-NN 相比，基于神经网络的方法可以更有效地提取图像对中的隐藏模式。

6.5.2　最大相关回归

为考察最大相关回归在实际应用中的性能，在深度学习领域常用的数据集开展相关实验，包括 MNIST[10], CIFAR-10[89], CIFAR-100[89] 等。为与经典的基于 Softmax 层与对数损失函数 (Log Loss) 的深度学习框架 (简记为 SL) 进行性能对比，构造经典学习框架的对照组 SL。对照组中特征提取的网络结构 $\mathbf{f}_\theta$ 与最大相关回归一致，但采用经典的 Softmax 层生成对后验概率 $P_{Y|X}$ 的估计，并通过最小化对数损失函数估计网络参数。

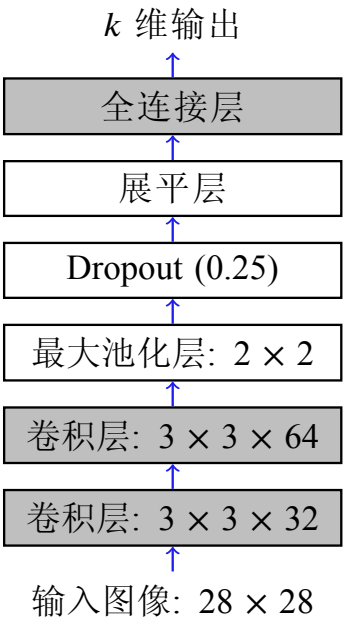


图 6.5　从 MNIST 数据集中提取特征的卷积神经网络结构，其中网络中的可训练参数用灰色的模块标出，Dropout 模块的细节可参考 [88] 。

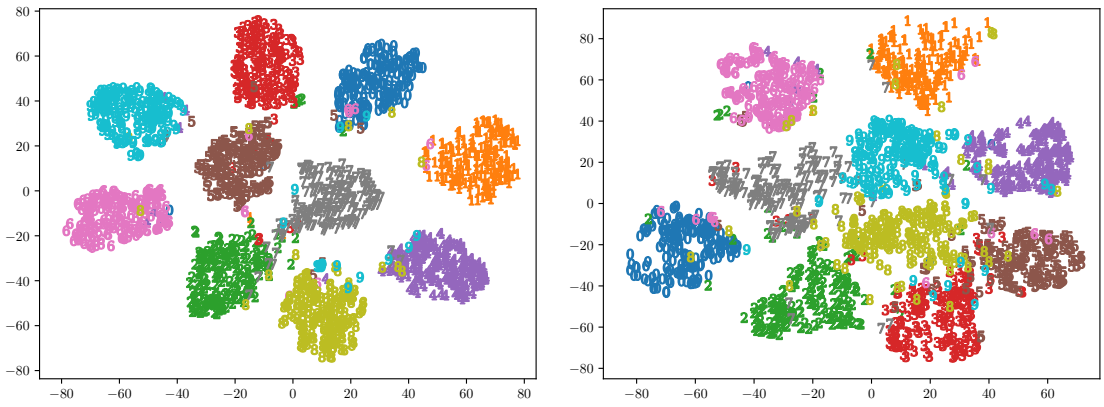
表 6.2　最大相关回归在 MNIST 数据集上测试准确率 (均值 ± 标准差) 与训练样本数的关系，及与经典 SL 神经网络上实验结果的对比。

训练样本数	最大相关回归 [%]	SL [%]
200	**82.11 ± 0.83**	77.75 ± 0.92
400	**90.73 ± 0.35**	84.70 ± 0.88
800	**94.77 ± 0.16**	91.83 ± 0.33
1000	**95.58 ± 0.31**	92.94 ± 0.38
1600	**96.45 ± 0.14**	94.80 ± 0.15
2000	**96.94 ± 0.17**	95.45 ± 0.18

6.5.2.1　MNIST 数据集

这里采用一个简易的两层卷积神经网络从 MNIST 数据集[10] 中提取特征，如图 6.5所示。实验中特征维度设为 $k = 10$。对最大相关回归与经典 SL 模型所对应的网络，均使用 ADADELTA[86] 训练 100 期，其中学习率设为 1.0 且衰减因子设为 0.95，小批量大小设为 32。当在训练集中所有 60,000 张图像上训练网络时，两个网络测试准确率均为 98.9%。进一步考察网络在小训练样本下的性能。具体而言，当只使用训练集前 n 个样本时，所得测试准确率随 n 的变化关系如表 6.2 所示，其中准确率的结果基于 10 次重复实验。可见与经典 SL 网络对比，最大相关回归可取得更好的性能，尤其是在小样本的场景下。

为深入理解该性能增益，考察当训练样本数 $n = 1000$ 时，分别由最大相关回



(a) MCR　　(b) SL

图 6.6　基于 $n = 1000$ 训练样本，由最大相关回归与经典 SL 网络在 MNIST 测试集上所提取特征 $\mathbf{f}_\theta$ 的可视化。

归及经典 SL 模型在测试集上所提取的特征 $\mathbf{f}_\theta$。具体地，使用 t-SNE[90] 将 10 维的 $\mathbf{f}_\theta$ 可视化到二维平面，如图 6.6 所示。该结果表明最大相关回归具有更大的类可分度，与第 6.3.4.2 节中相关讨论一致。

6.5.2.2　CIFAR-10

进一步地，在 CIFAR-10[89] 数据集上开展相关实验，该数据及包括 50,000 训练图片及 10,000 个测试图片，数据一共可分为 10 类。具体而言，用 ResNet-18[4] 从数据集中提取维度为 $k = 512$ 的特征 $\mathbf{f}_\theta$，并采用 [4] 中介绍的针对 SL 模型设计的训练设置如下：使用权重衰减为 5×10^{-4}、动量值为 0.9 的随机梯度下降法，分别训练最大相关回归及 SL 模型至 350 期。在训练过程中，学习率初始值设为 0.1 且分别在第 150 及 250 期减少为原值的 1/10。除此之外，这里采用如下的数据增强[4] 方法：首先将图像每边填充 4 个像素，并从该图像或其水平翻转图像中随机裁剪出 32×32 大小的图像。当小批量大小分别取 64、128、256 及 512 时，表 6.3 给出了最大相关回归与经典方法对应的测试准确率，其中均值与标准差为 10 次重复试验的结果。

此外，图 6.7 给出了当小批量大小取 128 时，两个方法在训练过程中的训练准确率及测试准确率。上述结果表明，在分类任务上最大相关回归可取得与经典方法相当的性能。值得指出的是，实验中并未针对最大相关回归方法微调训练超参数，而是直接使用在经典模型上调好的超参数，因此实践中最大相关回归的性能仍有一定的提升空间。

表 6.3　最大相关回归在 CIFAR-10 数据集上测试准确率 (均值 ± 标准差) 与小批量大小的关系，及与经典 SL 神经网络上实验结果的对比。

小批量大小	最大相关回归 [%]	SL [%]
64	94.30 ± 0.24	94.92 ± 0.16
128	94.87 ± 0.08	95.20 ± 0.11
256	95.02 ± 0.10	95.33 ± 0.12
512	94.56 ± 0.25	95.00 ± 0.15

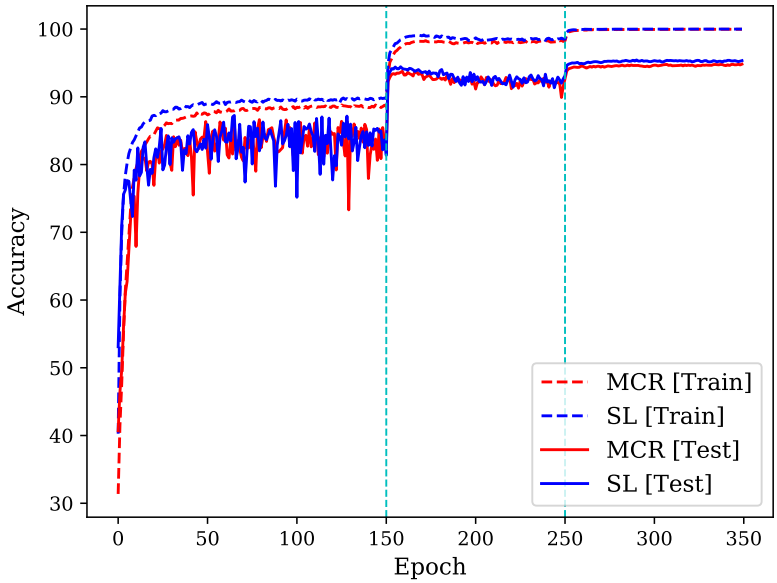


图 6.7　最大相关回归在 CIFAR-10 数据集上的测试与训练准确率及与经典 SL 模型的对比

6.5.2.3　CIFAR-100

基于与前述 CIFAR-10 实验类似的设置，进一步在 CIFAR-100[89] 数据集上研究算法的性能，该数据集共包含 100 类数据。这里同样采用 ResNet-18 网络由输入图像中提取 $k = 512$ 维的特征，并采用针对 SL 模型调节的训练参数。具体地，使用随机梯度下降法对最大相关回归模型与 SL 模型分别训练 200 期，其中权重衰减设为 5×10^{-4}，动量设为 0.9。在训练过程中，学习率初始值设为 0.1 且分别在第 60、120 及 160 期减少为原值的 1/5。在数据增强方法上，除采用 CIFAR-10 实验中的填充、裁剪及翻转操作外，还对图像进行了 $-15°$ 至 $15°$ 间的随机旋转。当小批量大小分别取 64、128、256 及 512 时，表 6.4 给出了最大相关回归与经典方法对应的测试准确率，其中均值与标准差为 10 次重复试验的结果。当小批量大小取 128 时，两个方法在训练过程中的训练准确率及测试准确率如图 6.8 所示。尽管未针对最大相关回归方法优化训练超参数，上述结果表明最大相关回归可以取得与经典

表 6.4　最大相关回归在 CIFAR-100 数据集上测试准确率 (均值 ± 标准差) 与小批量大小的关系，及与经典 SL 神经网络上实验结果的对比。

小批量大小	最大相关回归 [%]	SL [%]
64	74.16 ± 0.28	76.34 ± 0.22
128	75.11 ± 0.23	75.76 ± 0.16
256	75.15 ± 0.20	75.14 ± 0.28
512	74.40 ± 0.17	74.35 ± 0.20

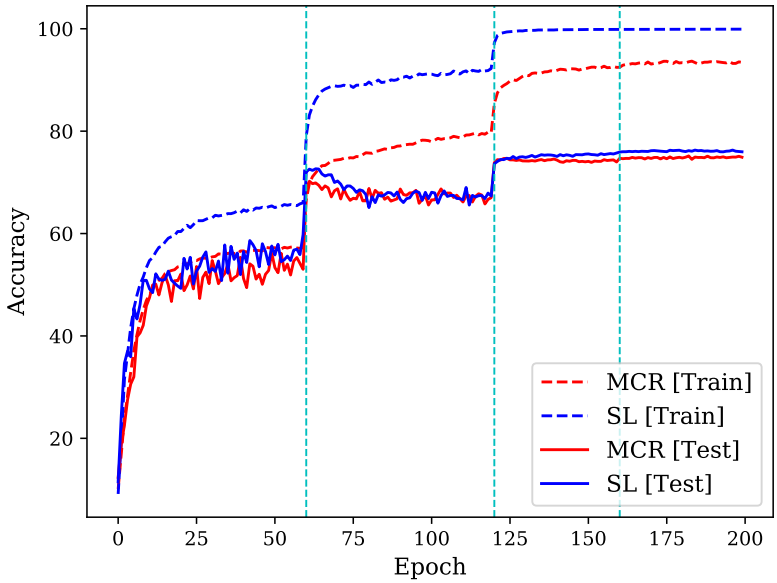


图 6.8　最大相关回归在 CIFAR-100 数据集上的测试与训练准确率及与经典 SL 模型的对比

SL 方法相当的性能。特别地，由表 6.4可知，当小批量大小设为 256 或 512 时，最大相关回归可以取得更好的性能；此外，由图 6.8 可知，最大相关回归测试准确率与训练准确率的差距相比经典方法更小，因此最大相关回归过拟合程度相对较轻。

6.5.3　无监督特征提取

为检验所设计的无监督特征提取算法在应用中的性能，这里在 MNIST[10] 手写体数据集上开展一系列实验。在 MNIST 数据集中，共有 $n = 60\,000$ 个训练图像，每个图像均由 28×28 的灰度值在 0 到 255 之间的像素构成，且与 “0” 到 “9” 之间的某个标签对应，以表示图像中手写体对应的数字。尽管该问题为有监督学习问题，实验结果表明算法 8 与第 6.4.2.3 节中讨论的低秩近似方法都可以在没有标签信息的情况下用于图像的特征提取，且这些以无监督方式提取的特征可在分类任务上取得良好的性能。

为明确 MNIST 识别问题中的诸随机变量 X_i，这里采用 [76] 中的划分方式，将

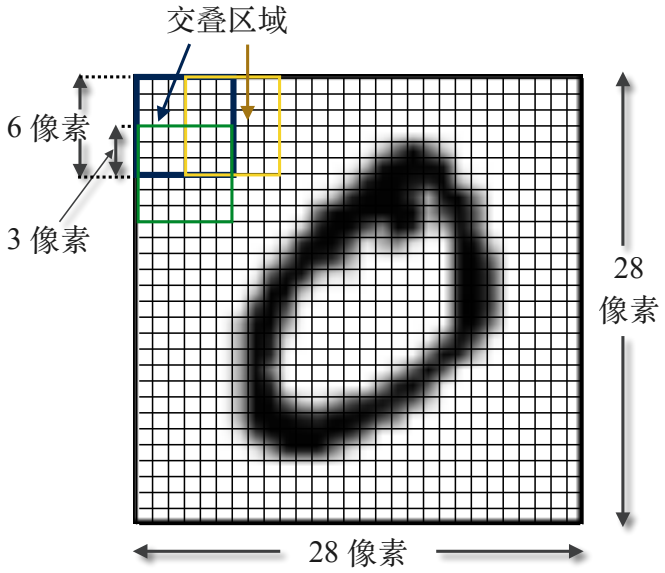


图 6.9　将 MNIST 图像划分为 $8\times 8=64$ 个 6×6 的子图像，使得相邻子图像有 3 像素的重合[76]。

每个图片样本分为 $8\times 8=64$ 个有交叠的 6×6 子图像，且每两个相邻子图像之间交叠 3 个像素，如 图 6.9 所示。通过该划分过程可避免直接考虑所有像素之间的相关性，在有效降低训练复杂度的同时获得相邻像素的相关性。于是第 i 个子图像可视为随机变量 X_i，$i=1,\cdots,64$。因此，若将 MNIST 数据集中第 ℓ 个样本的 i 个子图像记为 $x_i^{(\ell)}$，则每个随机变量 X_i 均对应 n 个训练样本 $x_i^{(1)},\cdots,x_i^{(n)}$。此外，我们将第 ℓ 个样本对应标签记为 $z_\ell\in\{0,\cdots,9\}$。

6.5.3.1　多元交替条件期望算法

为便于应用多元交替条件期望算法 (算法 8)，这里采用 [76] 中的图像预处理方式：首先，以 40 为阈值，将图像灰度量化为二元取值 “0” 和 “1”。经量化后，每个 $x_i^{(\ell)}$ 均为 36 维的二元向量，故字母集大小为 $|\mathcal{X}_i|=2^{36}$。为进一步降低字母集规模, 对每个子区域 i 遍历所有 n 个训练图像以得到所有 $\{0,1\}^{36}$ 中可取到的二元向量，再将这些二元向量映射到更小的字母集，使 Hamming 距离不超过 3 的两个二元向量划分到同一个字母集。该量化过程可总结为算法 9。

经预处理之后得 64 个随机变量 X_i，且第 ℓ 个图像对应于 64 维向量 $(x_1^{(\ell)},\cdots,x_{64}^{(\ell)})$, $\ell=1,\cdots,n$。应用算法 8 计算每个随机变量 X_i 对应的 k 个特征 $\vec{f}_i=(f_i^{(1)},\cdots,f_i^{(k)})$，该计算过程将第 ℓ 个预处理后的图像映射为 $64k$ 维的评分函数

$$\vec{s}_\ell=\left(\vec{f}_1(x_1^{(\ell)}),\cdots,\vec{f}_{64}(x_{64}^{(\ell)})\right),$$

算法 9 字母集归约的量化过程

Require: 训练样本 $\left\{x_i^{(\ell)}:\ \ell=1,\cdots,n\right\}$

初始化：令 $\mathcal{X}_i \leftarrow \varnothing$.
for $\ell = 1:n$ **do**
　if $\exists\, x \in \mathcal{X}_i$ 使得 $d_H\left(x, x_i^{(\ell)}\right) \leqslant 3$. **then**
　　$x_i^{(\ell)} \leftarrow x$.
　else
　　$\mathcal{X}_i \leftarrow \mathcal{X}_i \cup \left\{x_i^{(\ell)}\right\}$
　end if
end for

从而完成了图像的非线性特征提取。需要注意的是，该特征提取过程仅与图像样本有关，无需利用标签信息。

在计算得评分函数之后，接着训练线性支持向量机 (Support Vector Machine，SVM)[91] 利用所提取特征 $\vec{s}_\ell, \ell = 1,\cdots,n$ 对标签 z_ℓ 进行分类。为在测试集上检验该线性分类器的性能，对测试集图片应用相同的预处理并通过 $\vec{f}_i$ 将预处理图像映射为 $(64k)$ 维的评分向量，再用之前训练好的支持向量机预测相应的标签，所得错误率与 k 的关系如下表所示。

k	4	8	12	16	20	24
误差率 [%]	4.74	2.44	2.36	2.21	2.15	2.08

注意该方法使用*一层*具信息性的评分函数将图像映射到特征空间，再对映射所得特征应用线性分类器，可得到与含*两层* Sigmoid 函数的神经网络类似的性能 (由 3 层全连接网络所得错误率为 2.95%，其中两个隐层节点数分别 500 及 150 个[2,10])。需要指出的是，神经网络的方法需基于图像的标签信息进行特征提取，而这里使用的无监督方法基于子图像之间的结构相关性，只利用了图像本身的信息。因此，该结果实质上阐释了有关随机变量公共结构的信息在实践中的应用。

6.5.3.2　基于神经网络框架的函数表示求解

应用第 6.4.2.3 节的方法，首先搭建 64 个神经网络 $\mathtt{NN}_1,\cdots,\mathtt{NN}_{64}$ 用于生成图像的表示 $\underline{f}_1(X_1),\cdots,\underline{f}_{64}(X_{64})$，其中每个子网络 $\mathtt{NN}_i$ 包含两个卷积层，具体结构如图 6.10 所示。使用负的多元 H 评分 $-H\left(\underline{f}_1(X_1),\cdots,\underline{f}_{64}(X_{64})\right)$ 作为损失函数，可训练这 64 个自网络的参数并获得最优的函数表示。

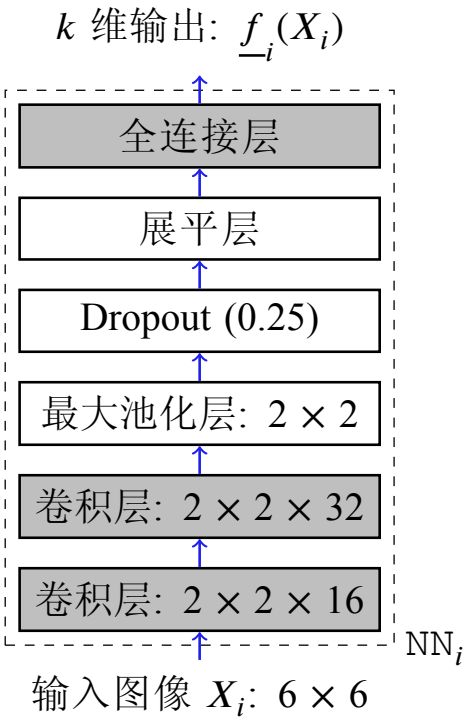


图 6.10　用于从输入 X_i 中提取特征 $\underline{f}_i(X_i)$ 的神经网络 NN_i 的结构

在训练集上所提取的函数表示基础上，训练线性支持向量机用于分类任务，并将训练好的支持向量机用于识别测试集，可得不同 k 对应的分类误差率如下所示。

k	4	8	12	16	20	24
误差率 [%]	3.46	1.73	1.43	1.17	1.15	1.11

与多元交替条件期望算法 (算法 8) 相比，基于神经网络的方法具有较明显的性能增益。产生该增益的主要原因是本方法直接使用卷积神经网络处理子图像，从而避免了前述方法中量化所带来的损失。需要指出的是，该无监督方法所提取特征可取得与基于卷积神经网络的有监督学习算法相似的性能。以经典的 LeNet-4[2] 为例，其在 MNIST 数据集上错误率为 1.1% 。

6.6　本章小结

基于局部信息几何分析方法，本章基于信息论准则设计不同场景下的机器学习算法，并给出了深度学习框架下的算法实现。具体而言，本章介绍了从数据样本计算最大相关函数的深度学习框架，在此基础上针对有监督学习中标签预测的问题，设计了最大相关回归算法。类似地，针对多变量的无监督特征提取的问题，由最小化变量间总相关的准则出发，设计了相应算法及基于深度学习框架的实现。理论分析及实际数据集上的测试结果都表明了所设计算法的有效性。

第 7 章　结论

7.1　工作归纳

伴随着机器学习算法，尤其是各类深度神经网络在工程实践中的广泛应用，高效、系统性的算法性能分析与设计的重要性不言而喻。然而，神经网络结构的多样性及实际数据自身的复杂性都为算法性能的系统性分析带来了挑战。本文基于局部信息几何的分析框架，从概率空间中的分布、有限维向量空间中信息向量及函数空间数据特征的一一对应关系出发，建立信息论、线性代数及机器学习三个不同层面问题的联系，并综合这三个问题分别在可解释性、分析及计算方面的优势。具体而言，本文的工作可归纳为以下三个部分：

首先，为解决神经网络可解释性问题，论文一方面从统计推断的通用特征选择问题出发，表明最优特征所对应信息向量为典型相关矩阵的奇异向量，从而与最大相关函数相对应；另一方面，从神经网络对数据特征的优化问题出发，通过引入信息向量表示，证明 Softmax 层输入特征对应信息向量的最优解也对应于典型相关矩阵的奇异向量，由此**建立神经网络所提取特征的信息论解释**。此外，对损失函数的分析表明，深度神经网络中对 Softmax 层的输入特征与权重的优化的数学本质为求解典型相关矩阵的低秩恢复问题。基于该低秩恢复结构，我们导出了特征与权重之间的交替更新算法，作为交替条件期望算法的多维推广；进一步地，根据该低秩恢复问题与损失函数的对应关系，我们**建立了评价特征性能的信息论度量**，即 H 评分函数；除此之外，由该低秩恢复问题中特征与权重之间的对称性，我们**证明了一般情况下神经网络特征与权重所满足的对称性**。

其次，为解决神经网络性能分析的问题，论文从信息向量的角度出发，分析了计算资源充分时泛化误差与训练样本数的关系，以及实际随机梯度下降的训练过程中平均泛化误差受超参数选择的影响。为刻画训练样本数对泛化误差的影响，基于 H 评分函数的性能度量，我们通过奇异向量的微扰分析计算泛化误差所对应的误差指数，并分别**求解了有监督学习及半监督学习中误差指数的解析表达式**。在此基础上，根据半监督学习的误差指数量化了有标签样本与无标签样本对学习任务的贡献，由此建立了总采样成本受限时对这两类样本的最优采样策略。在随机梯度下降的训练过程的分析中，我们的分析**表明了训练过程中计算效率与平均泛化误差所满足的折中关系**。进一步地，在大样本小学习率的假设下，我们**求解了平均泛化误差的解析表达式，从而量化了学习率与小批量大小对泛化误差的影响**，并给出了这两个超参数的最优取值。此外，我们量化分析的结果也可为包括线性

缩放准则在内的实用超参数调节技巧提供理论解释。

最后，为解决算法设计问题，论文基于信息论度量构建对概率分布的优化问题，以有限维空间的信息向量作为纽带，将信息论中概率空间的分布的优化问题转化为机器学习中对数据特征的给定泛函优化的问题，从而可设计基于深度学习的最优特征高效求解算法。基于该设计框架，我们**针对有监督学习及多模态无监督学习问题分别设计了最优特征提取的深度学习算法**，其中损失函数均可表示为 H 评分函数的形式。理论分析表明，所设计算法可分别解释为线性判别分析、主成分分析等经典的线性特征提取算法的非线性推广；数据集上一系列实验结果进一步检验了算法的有效性。

7.2　分析方法评注

作为本文的核心分析工具，局部信息几何方法建立了概率分布、向量空间与随机变量特征表示的一一对应关系，并进一步通过对 K-L 散度等关键信息度量的局部分析与优化，构建了机器学习问题与信息论问题间的联系。该方法的重要性体现在如下几个层面：

1. 在分析层面上，该方法导出了联合分布的特征模式分解，其数学上可表示为典型相关矩阵的奇异值分解。基于该模式分解，可将随机变量对间的相关性分解为彼此正交的信息，每部分可用典型相关矩阵的左右奇异向量表示，而信息的重要性对应相应的奇异值。
2. 在操作意义层面上，该方法建立了信息度量与有限维空间的联系，一方面为有限维空间信息向量的诸运算赋予操作意义，另一方面信息度量的优化过程分析可转化为有限维空间线性算子性质的分析，从而提供了对信息度量分析的有效途径。
3. 在方法层面，局部假设可视为一般问题的特例，因此在该假设下得到的结论可帮助理解问题的本质属性，并为一般情况下性质的分析提供启发。作为该分析模式的典型案例，第 3.6 节中对神经网络对称性的分析源于局部分析机制中的结论：在局部分析机制下，分析所揭示的神经网络特征提取的奇异值分解结构自然蕴含对称性，而一般情况下对称性则不显然。
4. 在算法设计层面，由局部假设给出的最优特征提取算法在一般情况下也具有良好的理论性质与实际性能，具体讨论可参考第 6 章。

由于上述特性，局部信息几何也可用于一般信息处理问题的分析，更详细的讨论可参考 [33]。

另外可注意到，局部分析方法依赖于相应假设的引入，如：随机变量间近似

独立的假设、或限制特征包含信息量较小的假设等 (参照定义 3.1、定义 3.2、及第 6.4 节中的低信息率假设等)。这类局部假设可类比于数学或物理中所考虑的无穷小量 (微元)，其引入本质是为了考察目标函数的局部性质。需要指出的是，正如数学中并不存在为无穷小量的数、真实世界中不存在可作为微元的物理量，实际应用往往并不满足近似独立或信息量较小的局部假设，但这不影响局部分析方法自身的有效性及重要性。

参考文献

[1] Bishop C M. Pattern recognition and machine learning. springer, 2006.

[2] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[3] LeCun Y, Bengio Y, Hinton G. Deep learning. nature, 2015, 521(7553):436-444.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] Orr G B, Müller K R. Neural networks: tricks of the trade. Springer, 2003.

[6] Goodfellow I, Bengio Y, Courville A, et al. Deep learning: volume 1. MIT press Cambridge, 2016.

[7] Shannon C E. A mathematical theory of communication. Bell system technical journal, 1948, 27(3):379-423.

[8] Gallager R G. Information theory and reliable communication: volume 2. Springer, 1968.

[9] Cover T M, Thomas J A. Elements of information theory. John Wiley & Sons, 2012.

[10] LeCun Y, Cortes C, Burges C. MNIST handwritten digit database [J/OL]. AT&T Labs, 1998. http://yann.lecun.com/exdb/mnist.

[11] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks // European conference on computer vision. Springer, 2014: 818-833.

[12] Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644, 2016.

[13] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle // 2015 IEEE Information Theory Workshop (ITW). IEEE, 2015: 1-5.

[14] Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810, 2017.

[15] Jakubovitz D, Giryes R, Rodrigues M R. Generalization error in deep learning // Compressed Sensing and Its Applications. Springer, 2019: 153-193.

[16] Bartlett P L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE transactions on Information Theory, 1998, 44(2):525-536.

[17] Friedman N, Yakhini Z. On the sample complexity of learning bayesian networks. arXiv preprint arXiv:1302.3579, 2013.

[18] Shamir O. Convergence of stochastic gradient descent for PCA // International Conference on Machine Learning. 2016: 257-265.

[19] Jain P, Jin C, Kakade S M, et al. Streaming PCA: Matching matrix bernstein and near-optimal finite sample guarantees for Oja’s algorithm // Conference on Learning Theory. 2016: 1147-1164.

[20] Allen-Zhu Z, Li Y. First efficient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate // Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on. IEEE, 2017: 487-492.

[21] Marinov T V, Mianjy P, Arora R. Streaming principal component analysis in noisy settings // International Conference on Machine Learning. 2018: 3410-3419.

[22] Li C J, Wang M, Liu H, et al. Near-optimal stochastic approximation for online principal component estimation. Mathematical Programming, 2018, 167(1):75-97.

[23] Li C J, Wang M, Liu H, et al. Diffusion approximations for online principal component estimation and global convergence // Advances in Neural Information Processing Systems. 2017: 645-655.

[24] LeCun Y, Bottou L, Orr G B, et al. Efficient backprop // Neural Networks: Tricks of the Trade. Springer, 1998: 9-50.

[25] Bengio Y. Practical recommendations for gradient-based training of deep architectures // Neural networks: Tricks of the trade. Springer, 2012: 437-478.

[26] Défossez A, Bach F. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions // Artificial Intelligence and Statistics. 2015: 205-213.

[27] Jain P, Kakade S M, Kidambi R, et al. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. 2018.

[28] Belghazi M I, Baratin A, Rajeswar S, et al. Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062, 2018.

[29] Ver Steeg G, Galstyan A. Discovering structure in high-dimensional data through correlation explanation // Advances in Neural Information Processing Systems. 2014: 577-585.

[30] Borade S, Zheng L. Euclidean information theory // 2008 IEEE International Zurich Seminar on Communications. IEEE, 2008: 14-17.

[31] Borade S, Zheng L. I-projection and the geometry of error exponents // Proceedings of the Forty-Fourth Annual Allerton Conference on Communication, Control, and Computing, Sept 27-29. 2006.

[32] Oja E. Simplified neuron model as a principal component analyzer. Journal of mathematical biology, 1982, 15(3):267-273.

[33] Huang S L, Makur A, Wornell G W, et al. On universal features for high-dimensional learning and inference. arXiv preprint arXiv:1911.09105, 2019.

[34] Huang S L, Xu X, Zheng L, et al. An information theoretic interpretation to deep neural networks // 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019: 1984-1988.

[35] Huang S L, Zheng L. Linear information coupling problems // 2012 IEEE International Symposium on Information Theory Proceedings. IEEE, 2012: 1029-1033.

[36] Huang S L, Makur A, Zheng L, et al. An information-theoretic approach to universal feature selection in high-dimensional inference // Information Theory (ISIT), 2017 IEEE International Symposium on. IEEE, 2017: 1336-1340.

[37] Hirschfeld H O. A connection between correlation and contingency. Proc. Cambridge Phil. Soc., 1935, 31:520-524.

[38] Gebelein H. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichungsrechnung. Z. für angewandte Math., Mech., 1941, 21: 364-379.

[39] Rényi A. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungarica, 1959, 10(3–4):441-451.

[40] Makur A, Kozynski F, Huang S L, et al. An efficient algorithm for information decomposition and extraction // Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on. IEEE, 2015: 972-979.

[41] Razaviyayn M, Farnia F, Tse D. Discrete Rényi Classifiers // Advances in Neural Information Processing Systems. 2015: 3276-3284.

[42] Breiman L, Friedman J H. Estimating optimal transformations for multiple regression and correlation. Journal of the American statistical Association, 1985, 80(391):580-598.

[43] Buja A, et al. Remarks on functional canonical variates, alternating least squares methods and ace. The Annals of Statistics, 1990, 18(3):1032-1069.

[44] Stoer J, Bulirsch R. Introduction to numerical analysis: volume 12. Springer Science & Business Media, 2013.

[45] Xu X, Huang S L, Zheng L, et al. The geometric structure of generalized softmax learning // 2018 IEEE Information Theory Workshop (ITW). IEEE, 2018: 1-5.

[46] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 1989, 2(4):303-314.

[47] Olga R, Jia D, Hao S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 2015, 115(3):211-252.

[48] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[49] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. // AAAI. 2017: 4278-4284.

[50] Chollet F. Xception: Deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357, 2016.

[51] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[52] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[53] Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks // Proceedings of the IEEE conference on computer vision and pattern recognition: volume 1. 2017: 3.

[54] Akaike H. Information theory and an extension of the maximum likelihood principle // Selected Papers of Hirotugu Akaike. Springer, 1998: 199-213.

[55] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer, 2009(8):30-37.

[56] Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika, 1936, 1(3):211-218.

[57] Huang S L, Xu X. On the sample complexity of HGR maximal correlation functions // 2019 IEEE Information Theory Workshop (ITW). 2019: 1-5.

[58] Csiszár I, Shields P C, et al. Information theory and statistics: A tutorial. Foundations and Trends® in Communications and Information Theory, 2004, 1(4):417-528.

[59] Xu X, Huang S L. On the asymptotic sample complexity of HGR maximal correlation functions in semi-supervised learning // 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2019: 879-886.

[60] Heath M T. Scientific computing: an introductory survey: volume 80. SIAM, 2018.

[61] Huang S L, Xu X. On the robustness of noisy ACE algorithm and multi-layer residual learning // 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019: 2474-2478.

[62] Xu X, Huang S L. On the optimal tradeoff between computational efficiency and generalizability of Oja’s algorithm. IEEE Access, 2020, 8:102616-102628.

[63] Horn R A, Johnson C R. Matrix analysis. 2nd ed. New York, NY, USA: Cambridge University Press, 2012.

[64] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. 2012: 1097-1105.

[65] Ge R, Kakade S M, Kidambi R, et al. The step decay schedule: A near optimal, geometrically decaying learning rate procedure. arXiv preprint arXiv:1904.12838, 2019.

[66] Vu V Q, Lei J, et al. Minimax sparse principal subspace estimation in high dimensions. The Annals of Statistics, 2013, 41(6):2905-2947.

[67] Wilson D R, Martinez T R. The general inefficiency of batch training for gradient descent learning. Neural Networks, 2003, 16(10):1429-1451.

[68] Hoffer E, Hubara I, Soudry D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks // Advances in Neural Information Processing Systems. 2017: 1731-1741.

[69] Smith S L, Kindermans P J, Ying C, et al. Don’t decay the learning rate, increase the batch size. arXiv preprint arXiv:1711.00489, 2017.

[70] Goyal P, Dollár P, Girshick R, et al. Accurate, large minibatch SGD: Training Imagenet in 1 hour. arXiv preprint arXiv:1706.02677, 2017.

[71] Wang L, Wu J, Huang S L, et al. An efficient approach to informative feature extraction from multimodal data // Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 5281-5288.

[72] Xu X, Huang S L. Maximal correlation regression. IEEE Access, 2020, 8:26591-26601.

[73] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition. Neural computation, 1989, 1(4):541-551.

[74] Fisher R A. The use of multiple measurements in taxonomic problems. Annals of eugenics, 1936, 7(2):179-188.

[75] Fukunaga K. Introduction to statistical pattern recognition (2nd ed.). San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[76] Huang S L, Zhang L, Zheng L. An information-theoretic approach to unsupervised feature selection for high-dimensional data // Information Theory Workshop (ITW), 2017 IEEE. IEEE, 2017: 434-438.

[77] Watanabe S. Information theoretical analysis of multivariate correlation. IBM Journal of research and development, 1960, 4(1):66-82.

[78] Huang S L, Xu X, Zheng L. An information-theoretic approach to unsupervised feature selection for high-dimensional data. IEEE Journal on Selected Areas in Information Theory, 2020.

[79] Jolliffe I T. Principal components in regression analysis [M/OL]. New York, NY: Springer New York, 1986: 129-155. https://doi.org/10.1007/978-1-4757-1904-8_8.

[80] Golub G H, Van Loan C F. Matrix computations: volume 3. Baltimore, Maryland, USA: Johns Hopkins University Press, 2013.

[81] Huang Q, Wang F, Guibas L. Functional map networks for analyzing and exploring large shape collections. ACM Trans. Graph., 2014, 33(4):36:1-36:11.

[82] Feizi S, Makhdoumi A, Duffy K, et al. Network maximal correlation. IEEE Transactions on Network Science and Engineering, 2017, 4(4):229-247.

[83] Feizi S, Tse D. Maximally correlated principal component analysis. arXiv preprint arXiv:1702.05471, 2017.

[84] Wang F, Huang Q, Guibas L J. Image co-segmentation via consistent functional maps // Proceedings of the 2013 IEEE International Conference on Computer Vision. Washington, DC, USA, 2013: 849-856.

[85] Wang F, Huang Q, Ovsjanikov M, et al. Unsupervised multi-class joint image segmentation // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC, USA, 2014: 3142-3149.

[86] Zeiler M D. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[87] Hsu H, Salamatian S, Calmon F P. Correspondence analysis using neural networks. arXiv preprint arXiv:1902.07828, 2019.

[88] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014, 15(1):1929-1958.

[89] Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. Citeseer, 2009.

[90] Maaten L v d, Hinton G. Visualizing data using t-SNE. Journal of machine learning research, 2008, 9(Nov):2579-2605.

[91] Cortes C, Vapnik V. Support-vector networks. Machine learning, 1995, 20(3):273-297.

[92] Dembo A, Zeitouni O. Large deviations techniques and applications. corrected reprint of the second (1998) edition. stochastic modelling and applied probability, 38. Springer-Verlag, Berlin, 2010.

[93] Polyanskiy Y, Wu Y. Lecture notes on information theory. Lecture Notes for ECE563 (UIUC) and, 2014, 6:2012-2016.

[94] Kato T. Perturbation theory for linear operators. 1976.

[95] Corless R M, Gonnet G H, Hare D E, et al. On the Lambert W function. Advances in Computational mathematics, 1996, 5(1):329-359.

[96] Magnus J R. On differentiating eigenvalues and eigenvectors. Econometric Theory, 1985, 1(2): 179-191.

致 谢

衷心感谢导师张林教授和黄绍伦助理教授对本人的精心指导，他们深厚的学术造诣和严谨治学风格将使我终生受益。

感谢麻省理工学院郑立中教授在课题上的指导与帮助，郑老师敏锐的洞察力为论文研究工作提供了具有深刻启发性的建议。

感谢清华电子工程系的王钺老师、袁坚老师，清华伯克利深圳学院的李阳老师、卡耐基梅隆大学的张旆老师和斯坦福大学的 Hae Young Noh 老师在工作展示方面提出的建议。

感谢清华–伯克利深圳学院物联网与社会物理信息系统实验室的同学们在博士工作期间提供的慷慨帮助，特别是李名扬、马飞、王立晨在实验设计与实现方面的支持，以及童鑫熠、王伟达对论文表达细节的改进建议。

感谢在博士期间帮助和鼓励过我的朋友们，包括陈鑫磊、关牧之、韩衍隽、黄星煜、黄延、姜蔚蔚、焦剑涛、孔德强、李炳霖、连婧、刘培钦、刘诗雨、刘心宇、刘轶铭、吕若辰、马蕊、唐子涵、徐素素、杨天宇、张博伦、张跃、郑亦欣等。

最后，我要感谢家人在博士工作期间给予的理解与支持。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

附录 A　第 3 章中的证明

A.1　定理 3.1 的证明

我们首先介绍与假设检验问题中误差指数相关的引理如下。

引理 A.1： 给定参考分布 $P_X \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}})$, 常数 $\epsilon > 0$ 及整数 n 与 k，令 $x_1, \cdots, x_n$ 表示由 P_1 或 P_2 生成的独立同分布样本，其中 $P_1, P_2 \in \mathcal{N}_\epsilon^{\mathcal{X}}(P_X)$。为判决这些样本由 P_1 还是 P_2 生成，构造 k 维统计量 $h^k = (h_1, \cdots, h_k)$ 使得

$$h_i = \frac{1}{n}\sum_{l=1}^{n} f_i(x_l), \quad i = 1, \cdots, k, \tag{A-1}$$

其中 $(f_1(X), \cdots, f_k(X))$ 为关于 P_X 零均值、单位方差且不相关的函数，亦即

$$\mathbb{E}_{P_X}\left[f_i(X)\right] = 0, \quad i \in \{1, \cdots, k\} \tag{A-2a}$$

$$\mathbb{E}_{P_X}\left[f_i(X) f_j(X)\right] = \delta_{ij}, \quad i, j \in \{1, \cdots, k\}. \tag{A-2b}$$

则当 $n \to \infty$ 时，基于 h^k 推断的误差概率随 n 指数衰减，对应的 (Chernoff) 误差指数为

$$\lim_{n\to\infty} \frac{-\log p_e}{n} \triangleq E_{h^k} = \sum_{i=1}^{k} E_{h_i}, \tag{A-3a}$$

其中

$$E_{h_i} = \frac{1}{8}\langle \phi_1 - \phi_2, \xi_i \rangle^2 + o(\epsilon^2), \tag{A-3b}$$

且 $\phi_1 \leftrightarrow P_1, \phi_2 \leftrightarrow P_2, \xi_i \leftrightarrow f_i(X), i \in \{1, \cdots, k\}$ 为相应的信息向量。

证明 注意到最优判决准则需将投影值

$$\sum_{i=1}^{k} h_i\left(\mathbb{E}_{P_1}\left[f_i(X)\right] - \mathbb{E}_{P_2}\left[f_i(X)\right]\right)$$

与某个阈值比较，由 Cramér 定理[92] 知真实分布为 P_j $(j = 1, 2)$ 时的误差指数为

$$E_j(\lambda) = \min_{P \in S(\lambda)} D(P \| P_j), \tag{A-4}$$

其中

$$\mathcal{S}(\lambda) \triangleq \left\{ P \in \mathcal{P}^{\mathcal{X}} : \mathbb{E}_P\left[f^k(X)\right] = \lambda\, \mathbb{E}_{P_1}\left[f^k(X)\right] + (1-\lambda)\, \mathbb{E}_{P_2}\left[f^k(X)\right] \right\}. \tag{A-5}$$

于是由 (A-2a) 可得

$$\begin{aligned}
\mathbb{E}_{P_j}\left[f_i(X)\right] &= \sum_{x\in\mathcal{X}} P_j(x)\, f_i(x) \\
&= \sum_{x\in\mathcal{X}} P_X(x)\, f_i(x) + \sum_{x\in\mathcal{X}} (P_j(x) - P_X(x)) f_i(x) \\
&= \mathbb{E}_{P_X}\left[f_i(X)\right] + \sum_{x\in\mathcal{X}} \sqrt{P_X(x)}\, \phi_j(x) \cdot \frac{\xi_i(x)}{\sqrt{P_X(x)}} \\
&= \sum_{x\in\mathcal{X}} \phi_j(x)\, \xi_i(x) \\
&= \langle \phi_j, \xi_i \rangle, \quad j = 1,2 \text{ 及 } i = 1, \cdots, k,
\end{aligned} \tag{A-6}$$

或等价表示为

$$\mathbb{E}_{P_j}\left[f^k(X)\right] = \langle \phi_j, \xi^k \rangle, \quad j = 1,2$$

其中 $\xi^k \triangleq (\xi_1, \cdots, \xi_k)$.

因此，可使用信息向量将约束 (A-5) 表示为

$$\langle \phi, \xi_i \rangle = \langle \lambda\, \phi_1 + (1-\lambda)\, \phi_2, \xi_i \rangle, \quad i = 1, \cdots, k,$$

亦即

$$\langle \phi, \xi^k \rangle = \langle \lambda\, \phi_1 + (1-\lambda)\, \phi_2, \xi^k \rangle. \tag{A-7}$$

令 P^* 表示 (A-4) 中最优的 P，则 P^* 属于通过 P_j 且自然统计量为 $f^k(x)$ 的指数族，即形如

$$\log \tilde{P}_{\theta^k}(x) = \sum_{i=1}^{k} \theta_i f_i(x) + \log P_j(x) - \alpha\left(\theta^k\right)$$

的 k 维分布族。注意到对任意分布 $P \in \mathcal{N}_\epsilon^{\mathcal{X}}(P_X)$ 及信息向量 $\phi \leftrightarrow P$，有

$$\begin{aligned}
\log P(x) &= \log P_X(x) + \log \frac{P(x)}{P_X(x)} \\
&= \log P_X(x) + \log\left(1 + \frac{1}{\sqrt{P_X(x)}} \phi(x)\right)
\end{aligned}$$

$$= \log P_X(x) + \frac{1}{\sqrt{P_X(x)}}\phi(x) + o(\epsilon),$$

故该指数族所对应信息向量为

$$\tilde{\phi}_{\theta^k}(x) = \sum_{i=1}^{k} \theta_i \xi_i(x) + \phi_j(x) - \alpha(\theta^k)\sqrt{P_X(x)} + o(\epsilon). \tag{A-8}$$

从而由 (A-2b) 可推出

$$\langle \tilde{\phi}_{\theta^k}, \xi_i \rangle = \theta_i + \langle \phi_j, \xi_i \rangle + o(\epsilon).$$

因此，根据 (A-7) 可知该指数族与线性族 (A-5) 交点为 $P^* = P_{\theta^{k*}}$，其中

$$\theta_i^* = \langle \lambda\phi_1 + (1-\lambda)\phi_2 - \phi_j, \xi_i \rangle + o(\epsilon),$$

故可得

$$\begin{aligned}
E_j(\lambda) &= D(P^* \| P_j) \\
&= \frac{1}{2}\left\| \tilde{\phi}_{\theta^k} - \phi_j \right\|^2 + o(\epsilon^2) && \text{(A-9a)} \\
&= \frac{1}{2}\left\| \sum_{i=1}^{k} \theta_i^* \xi_i \right\|^2 + \frac{1}{2}\alpha\left(\theta^{k*}\right)^2 + o(\epsilon^2) && \text{(A-9b)} \\
&= \frac{1}{2}\sum_{i=1}^{k} (\theta_i^*)^2 + \frac{1}{2}\alpha\left(\theta^{k*}\right)^2 + o(\epsilon^2) && \text{(A-9c)} \\
&= \frac{1}{2}\sum_{i=1}^{k} \langle \lambda\phi_1 + (1-\lambda)\phi_2 - \phi_j, \xi_i \rangle^2 + o(\epsilon^2), && \text{(A-9d)}
\end{aligned}$$

其中 (A-9a) 依据 K-L 散度的局部近似 (参考命题 2.1)，(A-9b) 基于 (A-8)，(A-9c) 基于 (A-2b)。为导出 (A-9d)，注意到由 $\theta^{k*} = O(\epsilon)$ 及

$$\alpha(0) = 0, \nabla\alpha(0) = \mathbb{E}_{P_j}\left[f^k(X)\right] = \langle \phi_j, \xi^k \rangle = O(\epsilon),$$

可得

$$\alpha\left(\theta^{k*}\right) = o(\epsilon^2).$$

最后，当 $\lambda = 1/2$ 有 $E_1(\lambda) = E_2(\lambda)$，故总误差概率对应的指数由 (A-3) 给出。 □

此外，证明中将用到 Markov 链中信息向量的如下性质。

引理 A.2： 给定 Markov 链 $X \leftrightarrow Y \leftrightarrow V$ 及 $v \in \mathcal{V}$，令 $\boldsymbol{\phi}_v^{X|V}$ 与 $\boldsymbol{\phi}_v^{Y|V}$ 表示 $P_{X|V}(\cdot|v)$ 及 $P_{Y|V}(\cdot|v)$ 的信息向量，则有

$$\boldsymbol{\phi}_v^{X|V} = \tilde{\mathbf{B}}^{\mathsf{T}} \boldsymbol{\phi}_v^{Y|V}. \tag{A-10}$$

证明 由 Markov 性可知

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y) P_Y(y),$$

$$\begin{aligned} P_{X|V}(x|v) &= \sum_{y \in \mathcal{Y}} P_{X|Y,V}(x|y,v) P_{Y|V}(y|v) \\ &= \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y) P_{Y|V}(y|v), \end{aligned}$$

因此有

$$P_{X|V}(x|v) - P_X(x) = \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y)[P_{Y|V}(y|v) - P_Y(y)].$$

相应的信息向量满足

$$\begin{aligned} \phi_v^{X|V}(x) &= \frac{1}{\sqrt{P_X(x)}} \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y) \sqrt{P_Y(y)} \phi_v^{Y|V}(y) \\ &= \sum_{y \in \mathcal{Y}} \left[\tilde{\mathbf{B}}(y,x) + \sqrt{P_X(x) P_Y(y)} \right] \phi_v^{Y|V}(y) \\ &= \sum_{y \in \mathcal{Y}} \tilde{\mathbf{B}}(y,x) \phi_v^{Y|V}(y), \end{aligned} \tag{A-11}$$

其中最后的等号基于

$$\sum_{y \in \mathcal{Y}} \sqrt{P_Y(y)} \phi_v^{Y|V}(y) = \sum_{y \in \mathcal{Y}} [P_{Y|V}(y|v) - P_Y(y)] = 0.$$

最后，将 (A-11) 表示为矩阵形式即得 (A-10)。 □

如下引理将用于计算 RIE 上的数学期望。

引理 A.3： 令 $\mathbf{z}$ 为球对称的 M 维随机向量，即对任意正交阵 $\mathbf{Q}$ 都有 $\mathbf{z} \overset{\mathrm{d}}{=} \mathbf{Q}\mathbf{z}$。若 $\mathbf{A}$ 为维度相容的给定矩阵，则

$$\mathbb{E}\left[\|\mathbf{z}^{\mathsf{T}} \mathbf{A}\|^2 \right] = \frac{1}{M} \mathbb{E}\left[\|\mathbf{z}\|^2 \right] \|\mathbf{A}\|_{\mathrm{F}}^2. \tag{A-12}$$

证明 由定义对任意正交阵 $\mathbf{Q}$ 有 $\mathbf{\Lambda_z} = \mathbf{Q}\mathbf{\Lambda_z}\mathbf{Q}^\mathsf{T}$，故 $\mathbf{\Lambda_z}$ 为对角阵。设 $\mathbf{\Lambda_z} = \lambda\,\mathbf{I}$, 则由

$$\operatorname{tr}\left\{\mathbf{\Lambda_z}\right\} = \mathbb{E}\left[\|\mathbf{z}\|^2\right] = \lambda M$$

可知

$$\lambda = \frac{1}{M}\operatorname{tr}\left\{\mathbf{\Lambda_z}\right\},$$

故

$$\mathbb{E}\left[\|\mathbf{z}^\mathsf{T}\mathbf{A}\|^2\right] = \operatorname{tr}\left\{\mathbf{A}^\mathsf{T}\mathbf{\Lambda_z}\mathbf{A}\right\} = \lambda\operatorname{tr}\left\{\mathbf{A}^\mathsf{T}\mathbf{A}\right\} = \frac{1}{M}\mathbb{E}\left[\|\mathbf{z}\|^2\right]\|\mathbf{A}\|_\mathrm{F}^2. \qquad \square$$

基于前述引理，定理 3.1 可证明如下。

证明 (定理 3.1 的证明) 由于判决结果只依赖于 f_i 的某个线性投影，可不妨假设 $\mathbb{E}_{P_X}\left[f_i(X)\right] = 0, i = 1, \cdots, k$。令 $\mathbf{f}$ 为 f^k 的向量表示，且记 $\tilde{\mathbf{f}} \triangleq \mathbf{\Lambda}_\mathbf{f}^{-1/2}\mathbf{f}$ 为归一化后的 $\mathbf{f}$，则相应的统计量 $\tilde{f}^k = (\tilde{f}_1, \cdots, \tilde{f}_k)$ 满足约束条件 (A-2)。构造统计量 $\tilde{h}^k = (\tilde{h}_1, \cdots, \tilde{h}_k)$ 为 [参见 (A-1)]

$$\tilde{h}_i = \frac{1}{n}\sum_{l=1}^{n}\tilde{f}_i(x_l), \quad i = 1, \cdots, k, \tag{A-13}$$

则由引理 A.1 可知基于 $\tilde{h}^k$ 区分 v 与 v' 的误差指数为

$$\begin{aligned} E_{\tilde{h}^k}(v, v') &= \frac{1}{8}\sum_{i=1}^{k}\left[\left(\boldsymbol{\phi}_v^{X|V} - \boldsymbol{\phi}_{v'}^{X|V}\right)^\mathsf{T}\tilde{\boldsymbol{\xi}}_i^X\right]^2 + o(\epsilon^2) \\ &= \frac{1}{8}\left\|\left(\boldsymbol{\phi}_v^{X|V} - \boldsymbol{\phi}_{v'}^{X|V}\right)^\mathsf{T}\tilde{\mathbf{\Xi}}^X\right\|^2 + o(\epsilon^2), \end{aligned}$$

其中 $\boldsymbol{\phi}_v^{X|V}$ 表示 $P_{X|V}(\cdot|v)$ 所对应的信息向量，$\tilde{\boldsymbol{\xi}}_i^X$ 表示 $\tilde{f}_i$ 的信息向量，且 $\tilde{\mathbf{\Xi}}^X \triangleq [\tilde{\boldsymbol{\xi}}_1^X, \cdots, \tilde{\boldsymbol{\xi}}_k^X]$。由于最优判决准则是线性的，对统计量进行线性变换不改变误差指数, 因此

$$\begin{aligned} E_{h^k}(v, v') &= E_{\tilde{h}^k}(v, v') \\ &= \frac{1}{8}\left\|\left(\boldsymbol{\phi}_v^{X|V} - \boldsymbol{\phi}_{v'}^{X|V}\right)^\mathsf{T}\tilde{\mathbf{\Xi}}^X\right\|^2 + o(\epsilon^2) \\ &= \frac{1}{8}\left\|\left(\boldsymbol{\phi}_v^{Y|V} - \boldsymbol{\phi}_{v'}^{Y|V}\right)^\mathsf{T}\tilde{\mathbf{B}}\tilde{\mathbf{\Xi}}^X\right\|^2 + o(\epsilon^2), \end{aligned} \tag{A-14}$$

其中在最后的等式中使用了引理 A.2 的结果。在 RIE 上取期望，并利用引理 A.3 的结论，可得

$$\mathbb{E}\left[E_{h^k}(v, v')\right] = \frac{1}{8}\mathbb{E}\left[\left\|\left(\boldsymbol{\phi}_v^{Y|V} - \boldsymbol{\phi}_{v'}^{Y|V}\right)^\mathsf{T}\tilde{\mathbf{B}}\tilde{\mathbf{\Xi}}^X\right\|^2\right] + o(\epsilon^2)$$

$$= \frac{\mathbb{E}\left[\left\|\boldsymbol{\phi}_v^{Y|V} - \boldsymbol{\phi}_{v'}^{Y|V}\right\|^2\right]}{8|\mathcal{Y}|}\left\|\tilde{\mathbf{B}}\tilde{\boldsymbol{\Xi}}^X\right\|_{\mathrm{F}}^2 + o(\epsilon^2),$$

又由 $\tilde{f}^k$ 定义有

$$\tilde{\boldsymbol{\Xi}}^X = \boldsymbol{\Xi}^X\left(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X\right)^{-\frac{1}{2}}, \qquad \square$$

由此得误差指数 (3-1)。

A.2　引理 3.1 的证明

首先给出两个相关的引理。

引理 A.4： 给定分布 $P \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}}), Q, R \in \mathcal{P}^{\mathcal{X}}$ 与充分小的 ϵ，设其满足 $D(P\|Q) \leqslant \epsilon^2$ 及 $D(P\|R) \leqslant \epsilon^2$，则存在独立于 ϵ 的常数 $C > 0$，使得 $D(Q\|R) \leqslant C\epsilon^2$。

证明 令 $\|\cdot\|_1$ 表示分布间的 ℓ_1 距离，亦即 $\|P - Q\|_1 \triangleq \sum_{x\in\mathcal{X}} |P(x) - Q(x)|$，则由 Pinsker's 不等式[9] 可得

$$\|P - Q\|_1 \leqslant \sqrt{2D(P\|Q)} < \sqrt{2}\epsilon, \tag{A-15a}$$

$$\|P - R\|_1 \leqslant \sqrt{2D(P\|R)} < \sqrt{2}\epsilon, \tag{A-15b}$$

从而有

$$\|Q - R\|_1 \leqslant \|P - Q\|_1 + \|P - R\|_1 \leqslant 2\sqrt{2}\epsilon. \tag{A-16}$$

此外，令 $p_{\min} \triangleq \min_{x\in\mathcal{X}} P(x)$，则对任意 $x \in \mathcal{X}$ 有

$$R(x) > P(x) - |P(x) - R(x)| \tag{A-17a}$$

$$> \min_{x\in\mathcal{X}} P(x) - \sqrt{2}\epsilon \tag{A-17b}$$

$$= p_{\min} - \sqrt{2}\epsilon, \tag{A-17c}$$

其中 (A-17b) 基于 (A-15b)。由于 $P \in \mathrm{relint}(\mathcal{P}^{\mathcal{X}})$，有 $p_{\min} > 0$，故对充分小的 ϵ 有 $R(x) > p_{\min}/2$。因此

$$D(Q\|R) \leqslant \sum_{x\in\mathcal{X}} \frac{(Q(x) - R(x))^2}{R(x)} \tag{A-18a}$$

$$\leqslant \frac{2}{p_{\min}} \sum_{x\in\mathcal{X}} [Q(x) - R(x)]^2 \tag{A-18b}$$

$$\leqslant \frac{2\|Q-R\|_1^2}{p_{\min}} \tag{A-18c}$$

$$\leqslant \frac{16}{p_{\min}}\epsilon^2, \tag{A-18d}$$

其中 (A-18a) 基于 K-L 散度的上界[93], (A-18d) 成立依据为 (A-16)。 □

引理 A.5： 对任意 $(x,y)\in\mathcal{X}\times\mathcal{Y}$，有

$$D(P_XP_Y\|P_X\,\tilde{P}_{Y|X}^{(g,b)})\geqslant P_X(x)\log\left[P_Y(y)e^{\tau(x,y)}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right]$$

其中 $\tilde{P}_{Y|X}^{(g,b)}$ 定义如 (3-2) 所示，且 $\tau(x,y)$ 定义为 $\tau(x,y)\triangleq\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)$。

证明 首先，可将条件分布 $\tilde{P}_{Y|X}^{(g,b)}(y|x)$ 表示为

$$\begin{aligned}\tilde{P}_{Y|X}^{(g,b)}(y|x)&=\frac{e^{g^{\mathsf{T}}(y)f(x)+b(y)}}{\sum_{y'\in\mathcal{Y}}e^{g^{\mathsf{T}}(y')f(x)+b(y')}}\\&=\frac{P_Y(y)e^{g^{\mathsf{T}}(y)f(x)+d(y)}}{\sum_{y'\in\mathcal{Y}}P_Y(y')e^{g^{\mathsf{T}}(y')f(x)+d(y')}}\\&=\frac{P_Y(y)e^{\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)}}{\sum_{y'\in\mathcal{Y}}P_Y(y')e^{\tilde{g}^{\mathsf{T}}(y')f(x)+\tilde{d}(y')}}\\&=\frac{P_Y(y)e^{\tau(x,y)}}{\sum_{y'\in\mathcal{Y}}P_Y(y')e^{\tau(x,y')}}.\end{aligned}\tag{A-19}$$

于是 K-L 散度 $D\big(P_XP_Y\big\|P_X\,\tilde{P}_{Y|X}^{(g,b)}\big)$ 可写作

$$\begin{aligned}D\big(P_XP_Y\big\|P_X\,\tilde{P}_{Y|X}^{(g,b)}\big)&=\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}P_X(x)P_Y(y)\log\frac{\sum_{y'\in\mathcal{Y}}P_Y(y)e^{\tau(x,y')}}{e^{\tau(x,y)}}\\&=\sum_{x\in\mathcal{X}}P_X(x)\log\left[\sum_{y'\in\mathcal{Y}}P_Y(y)e^{\tau(x,y')}\right]-\mathbb{E}_{P_XP_Y}\left[\tau(X,Y)\right]\\&=\sum_{x\in\mathcal{X}}P_X(x)\log\left[\sum_{y'\in\mathcal{Y}}P_Y(y)e^{\tau(x,y')}\right],\end{aligned}\tag{A-20}$$

其中最后的等号是依据 $\mathbb{E}_{P_XP_Y}\left[\tau(X,Y)\right]=0$。由此可得

$$\begin{aligned}D\big(P_XP_Y\big\|P_X\,\tilde{P}_{Y|X}^{(g,b)}\big)&\geqslant P_X(x)\log\left[\sum_{y'\in\mathcal{Y}}P_Y(y')e^{\tau(x,y')}\right]\\&\geqslant P_X(x)\log\left[P_Y(y)e^{\tau(x,y)}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right],\end{aligned}$$

其中第二个不等号依据 Jensen 不等式：

$$\begin{aligned}\sum_{y'\in\mathcal{Y}} P_Y(y')e^{\tau(x,y')} &= P_Y(y)e^{\tau(x,y)} + (1-P_Y(y))\sum_{y'\neq y}\frac{P_Y(y')}{1-P_Y(y)}e^{\tau(x,y')}\\ &\geqslant P_Y(y)e^{\tau(x,y)} + (1-P_Y(y))\exp\left(\frac{1}{1-P_Y(y)}\sum_{y'\neq y}P_Y(y')\tau(x,y')\right)\\ &= P_Y(y)e^{\tau(x,y)} + (1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}.\end{aligned}$$

□

基于前述结论，引理 3.1 可证明如下。

证明 (引理 3.1 的证明) 注意到当 $g=d=0$ 时，有 $\tilde{P}_{Y|X}^{(g,b)}=P_Y$，故 (3-4) 中最优的 g,d 满足

$$\begin{aligned}D(P_{XY}\|P_X\,\tilde{P}_{Y|X}^{(g,b)}) &\leqslant D(P_{XY}\|P_XP_Y)\\ &\leqslant \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\frac{\left[P_{XY}(x,y)-P_X(x)P_Y(y)\right]^2}{P_X(x)P_Y(y)}\\ &\leqslant \epsilon^2,\end{aligned}$$

其中第二个不等式依据 K-L 散度的上界[93]，最后的不等号根据的是 ϵ 相关的定义。

因为 $P_{XY}\in\text{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$，由引理 A.4 知存在 $C>0$ 及 $\epsilon_1>0$ 使得对任意 $\epsilon<\epsilon_1$ 有 $D(P_XP_Y\|P_X\,\tilde{P}_{Y|X}^{(g,b)})<C\epsilon^2$。此外，由引理 A.5 可知对所有 $(x,y)\in\mathcal{X}\times\mathcal{Y}$ 及 $\epsilon\in(0,\epsilon_1)$，有

$$C\epsilon^2 \geqslant P_X(x)\log\left[P_Y(y)e^{\tau(x,y)}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right]. \tag{A-21}$$

由于

$$\log\left[P_Y(y)e^{\tau(x,y)}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right]. = \frac{P_Y(y)}{2(1-P_Y(y))}\tau^2(x,y)+o(\tau^2(x,y)),$$

存在与 ϵ_1 取值无关的 $\delta>0$，使得对任意的 $|\tau(x,y)|\leqslant\delta$，都有

$$\log\left[P_Y(y)e^{\tau(x,y)}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right] > \frac{P_Y(y)}{2}\tau^2(x,y). \tag{A-22}$$

因此，若 $|\tau(x,y)|>\delta$，有

$$\begin{aligned}&\log\left[P_Y(y)e^{\tau(x,y)}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right]\\ &\geqslant \min\left\{\log\left[P_Y(y)e^{\delta}+(1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\delta}\right],\log\left[P_Y(y)e^{-\delta}+(1-P_Y(y))e^{\frac{P_Y(y)}{1-P_Y(y)}\delta}\right]\right\}\end{aligned}$$

$$\geqslant \frac{P_Y(y)}{2}\delta^2,$$

其中第二个不等式基于函数 $P_Y(y)e^t + (1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}t}$ 的单调性，第三个不等式基于 (A-22)。

故

$$\log\left[P_Y(y)e^{\tau(x,y)} + (1-P_Y(y))e^{-\frac{P_Y(y)}{1-P_Y(y)}\tau(x,y)}\right] > \frac{P_Y(y)}{2}\cdot\min\{\delta^2, \tau^2(x,y)\}, \quad \text{(A-23)}$$

从而由 (A-21) 得

$$C\epsilon^2 \geqslant \frac{P_X(x)P_Y(y)}{2}\cdot\min\{\delta^2, \tau^2(x,y)\}, \quad \text{(A-24)}$$

因此 $\tau(x,y) = O(\epsilon)$。实际上，令 $\epsilon_2 \triangleq \frac{\delta}{\sqrt{2C}}\cdot\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\sqrt{P_X(x)P_Y(y)}$, $\epsilon_0 \triangleq \min\{\epsilon_1, \epsilon_2\}$，则对任意 $\epsilon < \epsilon_0$，有

$$C\epsilon^2 < \frac{P_X(x)P_Y(y)}{2}\cdot\delta^2, \qquad \square$$

于是由 (A-24) 推出 $|\tau(x,y)| < C'\epsilon$，其中 $C' = \sqrt{\frac{2C}{P_X(x)P_Y(y)}}$。

A.3　引理 3.2 的证明

证明 由引理 3.1，存在 $C' > 0$ 使得对任意 $(x,y)\in\mathcal{X}\times\mathcal{Y}$，有

$$|\tilde{g}^{\mathsf{T}}(y)f(x) + \tilde{d}(y)| < C'\epsilon, \quad \text{(A-25)}$$

从而可得

$$|\mu_f^{\mathsf{T}}\tilde{g}(y) + \tilde{d}(y)| < C\epsilon, \quad \text{(A-26)}$$

$$|\tilde{g}^{\mathsf{T}}(y)\tilde{f}(x)| < 2C\epsilon, \quad \text{(A-27)}$$

其中 $C = \max\{C', 1\}$。

根据 (A-19) 的结论，可不妨假设 $\mathbb{E}_{P_Y}[g(Y)] = \mathbb{E}_{P_Y}[d(Y)] = 0$，则 (3-2) 可表示为

$$\tilde{P}_{Y|X}^{(g,b)}(y|x) = \frac{P_Y(y)e^{\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)}}{\sum_{y'\in\mathcal{Y}}P_Y(y')e^{\tilde{g}^{\mathsf{T}}(y')f(x)+\tilde{d}(y')}}. \quad \text{(A-28)}$$

故基于 (A-25) 的结果，(3-2) 的分母可近似为

$$\begin{aligned} P_Y(y)e^{\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)} &= P_Y(y)\left(1+\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)+o(\epsilon)\right) \\ &= P_Y(y)\left(1+\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)\right)+o(\epsilon). \end{aligned}$$

类似地，由

$$\begin{aligned} \sum_{y'\in\mathcal{Y}} P_Y(y)e^{\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)} &= \sum_{y'\in\mathcal{Y}} P_Y(y)\left(1+\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)\right)+o(\epsilon) \\ &= 1+\mathbb{E}_{P_Y}\left[\tilde{g}^{\mathsf{T}}(Y)f(x)\right]+\mathbb{E}_{P_Y}\left[\tilde{d}(y)\right]+o(\epsilon) \\ &= 1+o(\epsilon) \end{aligned}$$

可得

$$\left[\sum_{y'\in\mathcal{Y}} P_Y(y)e^{\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)}\right]^{-1} = [1+o(\epsilon)]^{-1} = 1+o(\epsilon).$$

故 (A-28) 可近似为

$$\begin{aligned} \tilde{P}^{(g,b)}_{Y|X}(y|x) &= \left[P_Y(y)\left(1+\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)\right)+o(\epsilon)\right][1+o(\epsilon)] && \text{(A-29)} \\ &= P_Y(y)\left(1+\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)\right)+o(\epsilon), && \text{(A-30)} \end{aligned}$$

因此对足够小的 ϵ 有 $P_X\,\tilde{P}^{(g,b)}_{Y|X} \in \mathcal{N}^{\mathcal{X}\times\mathcal{Y}}_{C\epsilon}(P_XP_Y)$。此外，基于分布的局部假设可得 $P_{XY} \in \mathcal{N}^{\mathcal{X}\times\mathcal{Y}}_{\epsilon}(P_XP_Y) \subset \mathcal{N}^{\mathcal{X}\times\mathcal{Y}}_{C\epsilon}(P_XP_Y)$，于是根据 K-L 散度的局部近似 (参考命题 2.1)

$$D(P_1\|P_2) = \frac{1}{2}\|\phi_1-\phi_2\|^2+o\left(\epsilon^2\right), \tag{A-31}$$

可得

$$\begin{aligned} D(P_{Y,X}\|P_X\,\tilde{P}^{(g,b)}_{Y|X}) &= \frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left[P_{Y,X}(y,x)-\tilde{P}^{(g,b)}_{Y|X}(y|x)P_X(x)\right]^2}{P_Y(y)P_X(x)}+o(\epsilon^2) \\ &= \frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\Bigg[\frac{P_{Y,X}(y,x)}{\sqrt{P_Y(y)P_X(x)}}-\sqrt{P_Y(y)P_X(x)} \\ &\quad -\sqrt{P_Y(y)P_X(x)}\left(\tilde{g}^{\mathsf{T}}(y)f(x)+\tilde{d}(y)+o(\epsilon)\right)\Bigg]^2+o(\epsilon^2) \\ &= \frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\Big[\tilde{\mathbf{B}}(y,x)-\sqrt{P_Y(y)P_X(x)}\tilde{g}^{\mathsf{T}}(y)\tilde{f}(x) \end{aligned}$$

$$
\begin{aligned}
&-\sqrt{P_Y(y)P_X(x)}\left(\tilde{d}(y)+\mu_f^{\mathsf{T}}\tilde{g}(y)\right)\\
&\quad-\sqrt{P_Y(y)P_X(x)}o(\epsilon)\Big]^2+o(\epsilon^2)\\
\overset{(*)}{=}&\frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left[\tilde{\mathbf{B}}(y,x)-\sqrt{P_Y(y)P_X(x)}\tilde{g}^{\mathsf{T}}(y)\tilde{f}(x)\right]^2\\
&+\frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left[\sqrt{P_Y(y)P_X(x)}\left(\tilde{d}(y)+\mu_f^{\mathsf{T}}\tilde{g}(y)\right)\right]^2+o(\epsilon^2)\\
=&\frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left[\tilde{\mathbf{B}}(y,x)-\left(\xi^Y(y)\right)^{\mathsf{T}}\xi^X(x)\right]^2\\
&+\frac{1}{2}\mathbb{E}_{P_Y}\left[(\tilde{d}(y)+\mu_f^{\mathsf{T}}\tilde{g}(y))^2\right]+o(\epsilon^2)\\
=&\frac{1}{2}\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\|_{\mathrm{F}}^2+\frac{1}{2}\gamma^{(g,b)}(f)+o(\epsilon^2),
\end{aligned}
$$

其中 $(*)$ 依据的是 (A-26)–(A-27)、$|\tilde{B}(y,x)|<\epsilon$、以及 (因 $\mathbb{E}\left[\tilde{d}(Y)\right]=0,\mathbb{E}\left[\tilde{f}(X)\right]=\mathbb{E}\left[\tilde{g}(Y)\right]=0$)

$$
\begin{aligned}
&\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\tilde{\mathbf{B}}(y,x)\sqrt{P_Y(y)P_X(x)}\left(\tilde{d}(y)+\mu_f^{\mathsf{T}}\tilde{g}(y)\right)=0,\\
&\sum_{x\in\mathcal{X},y\in\mathcal{Y}}P_Y(y)P_X(x)\tilde{g}^{\mathsf{T}}(y)\tilde{f}(x)\left(\tilde{d}(y)+\mu_f^{\mathsf{T}}\tilde{g}(y)\right)=0.
\end{aligned}
$$

□

A.4　定理 3.2 与定理 3.3 的证明

基于引理 3.2 的结论，可得 定理 3.2 与定理 3.3 的证明。

证明 (定理 3.2 与定理 3.3 的证明) 由于 $d(\cdot)$ 取值仅影响 K-L 散度表达式中的第二项，总可以取 $d(\cdot)$ 使得 $\tilde{d}(y)+\mu_f^{\mathsf{T}}\tilde{g}(y)=0$，则最优的 $(\boldsymbol{\Xi}^Y,\boldsymbol{\Xi}^X)$ 满足

$$
(\boldsymbol{\Xi}^Y,\boldsymbol{\Xi}^X)^*=\arg\min_{(\boldsymbol{\Xi}^Y,\boldsymbol{\Xi}^X)}\left\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2. \tag{A-32}
$$

令导数[①]

$$
\frac{\partial}{\partial\boldsymbol{\Xi}^Y}\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\|_{\mathrm{F}}^2=2(\boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X-\tilde{\mathbf{B}}\boldsymbol{\Xi}^X) \tag{A-33}
$$

等于零，可得对给定 $\boldsymbol{\Xi}^X$ 最优的 $\boldsymbol{\Xi}^Y$ 为 [②]

$$
\boldsymbol{\Xi}^{Y*}=\tilde{\mathbf{B}}\boldsymbol{\Xi}^X(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X)^{-1}. \tag{A-34}
$$

① 这里约定标量对矩阵导数与矩阵自身维度相同，即采用矩阵微积分中的分母记法。

② 此处假设 $\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X$，亦即 $\boldsymbol{\Lambda}_{\tilde{f}(X)}$ 可逆。若 $\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\boldsymbol{\Xi}^X$ 奇异，需将结论中的矩阵逆改为其 Moore–Penrose 逆 (伪逆)。

由 $\mathbf{1}^{\mathsf{T}}\sqrt{\mathbf{P}_Y}\,\tilde{\mathbf{B}}=0$ 可知 $\mathbf{1}^{\mathsf{T}}\sqrt{\mathbf{P}_Y}\,\boldsymbol{\Xi}^{Y*}=0$，故 $\boldsymbol{\Xi}^{Y*}$ 所对应的特征函数均值为零。

为将 (A-34) 中的 $\boldsymbol{\Xi}^{Y*}$ 用 f 及 g 表示，注意到由 命题 2.3 可得 $\mathbb{E}_{P_{X|Y}}[f(X)|Y]\leftrightarrow\tilde{\mathbf{B}}\boldsymbol{\phi}$，同理 $\boldsymbol{\Lambda}_{\tilde{f}(X)}=\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\boldsymbol{\Xi}^{X}$，故 (A-34) 等价于

$$\tilde{g}^{*}(y)=\mathbb{E}_{P_{X|Y}}\left[\boldsymbol{\Lambda}_{\tilde{f}(X)}^{-1}\tilde{f}(X)\,\middle|\,Y=y\right]. \tag{A-35}$$

由对称性可立即得 定理 3.3 的前两个等式。为证明两个定理中的第三个等式，需最小化 $\gamma^{(g,b)}(f)=\mathbb{E}_{P_Y}\left[(\mu_f^{\mathsf{T}}\tilde{g}(Y)+\tilde{d}(Y))^2\right]$。对给定 $\tilde{g}$ 及 μ_f，最优的 $\tilde{d}$ 为

$$\tilde{d}^{*}(y)=-\mu_f^{\mathsf{T}}\tilde{g}(Y), \tag{A-36}$$

其对应的 $\gamma^{(g,b)}(f)=0$.

当固定 $\tilde{d}$ 与 $\tilde{g}$ 时，可得

$$\begin{aligned}\gamma^{(g,b)}(f)&=\mathbb{E}_{P_Y}\left[(\mu_f^{\mathsf{T}}\tilde{g}(Y)+\tilde{d}(Y))^2\right]\\&=\mu_f^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\mu_f+2\mu_f^{\mathsf{T}}\mathbb{E}_{P_Y}\left[\tilde{g}(Y)\tilde{d}(Y)\right]+\operatorname{var}(\tilde{d}(Y)).\end{aligned} \tag{A-37}$$

令 $\frac{\partial}{\partial\mu_f}\gamma^{(g,b)}(f)=0$ 可得

$$\mu_f^{*}=-\boldsymbol{\Lambda}_{\tilde{g}(Y)}^{-1}\mathbb{E}_{P_Y}\left[\tilde{g}(Y)\tilde{d}(Y)\right]. \tag{A-38}$$

□

A.5　定理 3.4 的证明

证明 由引理 3.2 可知选取最优的 $(\boldsymbol{\Xi}^{Y},\boldsymbol{\Xi}^{X})$ 等价于求解 $\tilde{\mathbf{B}}$ 的矩阵分解问题。由于 $\boldsymbol{\Xi}^{Y}$ 与 $\boldsymbol{\Xi}^{X}$ 的秩均不超过 k，由 Eckart–Young–Mirsky 定理[56] 可知，最优的 $\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}$ 为 $\tilde{\mathbf{B}}$ 保留前 k 个模式的截断奇异值分解 (Truncated Singular Value Decomposition)。故 $(\boldsymbol{\Xi}^{Y},\boldsymbol{\Xi}^{X})^{*}$ 分别为对应于 $\tilde{\mathbf{B}}$ 前 k 个奇异值的左奇异向量与右奇异向量。

偏置项 $\tilde{d}(y)=-\mu_f^{\mathsf{T}}\tilde{g}(y)$ 的最优性已于附录 A.4 中证明。

A.6　定理 3.5 的证明

首先给出如下引理。

引理 A.6 (Pythagorean 定理)： 给定 $\boldsymbol{\Xi}^{Y}$，令 $\boldsymbol{\Xi}^{X*}$ 表示由 (3-9) 定义的最优的 $\boldsymbol{\Xi}^{X}$，则

$$\left\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}-\left\|\tilde{\mathbf{B}}-\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}=\left\|\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}-\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}. \tag{A-39}$$

证明 记 $\langle \mathbf{U}, \mathbf{V} \rangle_{\mathsf{F}}$ 为矩阵 $\mathbf{U}$ 及 $\mathbf{V}$ 的 Frobenius 内积，即 $\langle \mathbf{U}, \mathbf{V} \rangle_{\mathsf{F}} \triangleq \mathrm{tr}\left\{ \mathbf{U}^{\mathsf{T}} \mathbf{V} \right\}$，则

$$\begin{aligned}
\left\langle \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}, \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\rangle_{\mathsf{F}} &= \mathrm{tr}\left\{ \tilde{\mathbf{B}} \boldsymbol{\Xi}^{X} \left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}} \right\} - \mathrm{tr}\left\{ \boldsymbol{\Xi}^{X*} \left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}} \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\} \\
&= \mathrm{tr}\left\{ \tilde{\mathbf{B}} \boldsymbol{\Xi}^{X} \left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}} \right\} - \mathrm{tr}\left\{ \tilde{\mathbf{B}}^{\mathsf{T}} \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\} \\
&= 0,
\end{aligned}$$

故

$$\begin{aligned}
\left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2} &= \left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} + \left(\boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right) \right\|_{\mathrm{F}}^{2} \\
&= \left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} \right\|_{\mathrm{F}} + \left\| \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2} \\
&\quad + 2 \left\langle \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}, \boldsymbol{\Xi}^{Y} \left(\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} - \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right) \right\rangle_{\mathsf{F}} \\
&= \left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} \right\|_{\mathrm{F}} + \left\| \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2}. \qquad \square
\end{aligned}$$

证明 (定理 3.5 的证明) 由引理 A.6 可知

$$\begin{aligned}
L(f) - L(f^{*}) &= \frac{1}{2} \left[\left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2} - \left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2} \right] \\
&\quad + \frac{1}{2} \left[\gamma^{(g,b)}(s) - \gamma^{(g,b)}(s^{*}) \right] + o(\epsilon^{2}) \\
&= \frac{1}{2} \left\| \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2} + \frac{1}{2} \kappa^{(g,b)}(f, f^{*}) + o(\epsilon^{2}),
\end{aligned}$$

其中 $\kappa^{(g,b)}(f, f^{*}) \triangleq \gamma^{(g,b)}(s) - \gamma^{(g,b)}(s^{*})$。接下来分别对 $\left\| \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}} - \boldsymbol{\Xi}^{Y} \left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^{2}$ 及 $\kappa^{(g,b)}(f, f^{*})$ 两项进行优化。

首先考察第一项。为此，将 $\boldsymbol{\Xi}^{X}$ 表达为 $\mathbf{W}$ 及 $\boldsymbol{\Xi}_{1}^{X}$ 的形式。由 (3-14) 可得

$$\mathbb{E}\left[f_{z}(X) \right] = \sigma\left(c(z) \right) + o(\epsilon), \tag{A-40a}$$

$$\tilde{f}_{z}(x) = w^{\mathsf{T}}(z) \tilde{t}(x) \cdot \sigma'\left(c(z) \right) + o(\epsilon), \tag{A-40b}$$

其可用信息向量表达为

$$\boldsymbol{\Xi}^{X} = \boldsymbol{\Xi}_{1}^{X} \mathbf{W}^{\mathsf{T}} \mathbf{J} + o(\epsilon). \tag{A-41}$$

又由定理 3.3 知

$$\boldsymbol{\Xi}^{X*} = \tilde{\mathbf{B}}^{\mathsf{T}} \boldsymbol{\Xi}^{Y} \left(\left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}} \boldsymbol{\Xi}^{Y} \right)^{-1}, \tag{A-42}$$

故

$$\begin{aligned}
\left\|\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}-\boldsymbol{\Xi}^{Y}\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^{2} &= \left\|\left(\left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}}\boldsymbol{\Xi}^{Y}\right)^{1/2}\left(\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}-\left(\boldsymbol{\Xi}^{X}\right)^{\mathsf{T}}\right)\right\|_{\mathrm{F}}^{2} \\
&= \left\|\left(\left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}}\boldsymbol{\Xi}^{Y}\right)^{1/2}\cdot\left(\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}-\mathbf{J}\mathbf{W}\left(\boldsymbol{\Xi}_{1}^{X}\right)^{\mathsf{T}}-o(\epsilon)\right)\right\|_{\mathrm{F}}^{2} \\
&= \left\|\left(\left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}}\boldsymbol{\Xi}^{Y}\right)^{1/2}\cdot\left(\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}-\mathbf{J}\mathbf{W}\left(\boldsymbol{\Xi}_{1}^{X}\right)^{\mathsf{T}}\right)\right\|_{\mathrm{F}}^{2}+o(\epsilon^{2}) \\
&= \left\|\left(\left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}}\boldsymbol{\Xi}^{Y}\right)^{1/2}\mathbf{J}\cdot\left(\mathbf{J}^{-1}\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}-\mathbf{W}\left(\boldsymbol{\Xi}_{1}^{X}\right)^{\mathsf{T}}\right)\right\|_{\mathrm{F}}^{2}+o(\epsilon^{2}) \\
&= \left\|\Theta\tilde{\mathbf{B}}_{1}-\Theta\mathbf{W}\left(\boldsymbol{\Xi}_{1}^{X}\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^{2}+o(\epsilon^{2}), \qquad \text{(A-43)}
\end{aligned}$$

其中第三个等号依据 [参见 (A-27)] $\tilde{f}(x)=O(\epsilon)$ 及 $\tilde{g}(y)=O(1)$，最后的等号基于定义 $\tilde{\mathbf{B}}_{1}\triangleq\mathbf{J}^{-1}\left(\boldsymbol{\Xi}^{X*}\right)^{\mathsf{T}}$ 和 $\Theta\triangleq\left(\left(\boldsymbol{\Xi}^{Y}\right)^{\mathsf{T}}\boldsymbol{\Xi}^{Y}\right)^{1/2}\mathbf{J}$。

对第二项，由 (A-37) 及 (A-38) 可知

$$\begin{aligned}
\kappa^{(g,b)}(f,f^{*}) &= [(\mu_{f}-\mu_{f^{*}})+\mu_{f^{*}}]^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\left[(\mu_{f}-\mu_{f^{*}})+\mu_{f^{*}}\right] \\
&\quad -\mu_{f^{*}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\mu_{f^{*}}+2(\mu_{f}-\mu_{f^{*}})^{\mathsf{T}}\mathbb{E}_{P_{Y}}\left[\tilde{g}(Y)\tilde{d}(Y)\right] \\
&= (\mu_{f}-\mu_{f^{*}})^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\left(\mu_{f}-\mu_{f^{*}}\right) \\
&\quad +2(\mu_{f}-\mu_{f^{*}})^{\mathsf{T}}\left(\boldsymbol{\Lambda}_{\tilde{g}(Y)}\mu_{f^{*}}+\mathbb{E}_{P_{Y}}\left[\tilde{g}(Y)\tilde{d}(Y)\right]\right) \\
&= (\mu_{f}-\mu_{f^{*}})^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\left(\mu_{f}-\mu_{f^{*}}\right). \qquad \text{(A-44)}
\end{aligned}$$

结合 (A-43) 及 (A-44) 的结果即得欲证结论。 □

接着，通过最小化 L 可求得 μ_{f}^{*} 与 w^{*}。与前述方法类似，这两项可独立求解。为求解 μ_{f}^{*}, 考虑隐层激活函数有界的情形，即 $\sigma_{\min}\preceq\mu_{f}\preceq\sigma_{\max}$，例如 sigmoid 函数 $1/(1+e^{-x})$ 或 $\tanh(x)$。则最优的 μ_{f} 为以下优化问题的解：

$$\begin{aligned}
&\underset{\mu_{f}}{\text{minimize}} \quad (\mu_{f}-\mu_{f^{*}})^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{g}(Y)}\left(\mu_{f}-\mu_{f^{*}}\right) \\
&\text{subject to} \quad \sigma_{\min}\preceq\mu_{f}\preceq\sigma_{\max}.
\end{aligned} \qquad \text{(A-45)}$$

若 $\mu_{f^{*}}$ 满足约束 (A-45)，则其为最优解；否则 μ_{f}^{*} 中某些元素将取到 $\sigma_{\min}$ 或 $\sigma_{\max}$，对应隐层节点的饱和现象。

此外，由 (A-40a) 知隐藏层的偏置项 $c(z)$ 为 ①

$$c(z)=\sigma^{-1}(\mu_{f}^{*}(z))+o(\epsilon).$$

① 若 $\mu_{t}\neq 0$，公式应修正为 $c(z)=\sigma^{-1}(\mu_{f}^{*}(z))-\mu_{t}^{\mathsf{T}}w+o(\epsilon)$.

为求解 $\mathbf{W}^*$，令

$$\begin{aligned}\tilde{\mathbf{B}}_1' &\triangleq \Theta\tilde{\mathbf{B}}_1 = \left(\left(\boldsymbol{\Xi}^Y\right)^{\mathsf{T}}\boldsymbol{\Xi}^Y\right)^{-1/2}\left(\boldsymbol{\Xi}^Y\right)^{\mathsf{T}}\tilde{\mathbf{B}},\\ \mathbf{W}' &\triangleq \Theta\mathbf{W} = \left(\left(\boldsymbol{\Xi}^Y\right)^{\mathsf{T}}\boldsymbol{\Xi}^Y\right)^{1/2}\mathbf{J}\mathbf{W},\end{aligned} \tag{A-46}$$

则最优的 $\mathbf{W}'$ 为优化问题

$$\underset{\mathbf{W}'}{\text{minimize}} \quad \|\tilde{\mathbf{B}}_1{}' - \mathbf{W}'\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\|_{\mathrm{F}}^2, \tag{A-47}$$

的解，从而可得

$$\mathbf{W}'^* = \tilde{\mathbf{B}}_1{}'\boldsymbol{\Xi}_1^X(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X)^{-1}. \tag{A-48}$$

因此

$$\begin{aligned}\mathbf{W}^* &= \Theta^{-1}\mathbf{W}'^* = \Theta^{-1}\tilde{\mathbf{B}}_1'\boldsymbol{\Xi}_1^X(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X)^{-1}\\ &= \tilde{\mathbf{B}}_1\boldsymbol{\Xi}_1^X(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X)^{-1}\\ &= \mathbf{J}^{-1}\cdot[\boldsymbol{\Xi}^Y(\left(\boldsymbol{\Xi}^Y\right)^{\mathsf{T}}\boldsymbol{\Xi}^Y)^{-1}]^{\mathsf{T}}\tilde{\mathbf{B}}\,\boldsymbol{\Xi}_1^X(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X)^{-1},\end{aligned}$$

其中 $\tilde{\mathbf{B}}\,\boldsymbol{\Xi}_1^X(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X)^{-1}$ 项对应于 $\tilde{t}(X)$ 的特征投影：

$$\tilde{\mathbf{B}}\,\boldsymbol{\Xi}_1^X\left(\left(\boldsymbol{\Xi}_1^X\right)^{\mathsf{T}}\boldsymbol{\Xi}_1^X\right)^{-1} \leftrightarrow \mathbb{E}_{P_{X|Y}}\left[\boldsymbol{\Lambda}_{\tilde{t}(X)}^{-1}\tilde{t}(X)\,\middle|\,Y\right]. \tag{A-49}$$

故多层神经网络可视作为计算不同层所提取特征的某种广义特征投影。注意到投影后的特征 $\mathbb{E}_{P_{\tilde{t}|Y}}\left[\boldsymbol{\Lambda}_{\tilde{t}}^{-1}\tilde{t}\middle|Y\right]$ 仅取决于分布 $P_{\tilde{t}|Y}$，不依赖于 $P_{X|Y}$，故在实际问题中进行该特征投影操作无需假定任何隐变量 X 的信息。

A.7　引理 3.3 的证明

证明 因 $P_{XY}\in\operatorname{relint}(\mathcal{P}^{\mathcal{X}\times\mathcal{Y}})$，$\exists\,\delta_P>0$ 使得 $\forall(x,y)\in\mathcal{X}\times\mathcal{Y}$, $P_{XY}(x,y)>\delta_P$。此外，可验证 $\exists\,\delta_Q>0$ 使得对任意满足 $D(P_{XY}\|Q_{XY})\leqslant D(P_{XY}\|P_XP_Y)$ 的 $Q_{XY}\in\mathcal{E}_k$ 有

$$\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} Q_{XY}(x,y)>\delta_Q.$$

事实上，对任意 $(x,y)\in\mathcal{X}\times\mathcal{Y}$，记 $p=P_{XY}(x,y), q=Q_{XY}(x,y)$，则

$$\begin{aligned}D(P_{XY}\|P_XP_Y) &\geqslant D(P_{XY}\|Q_{XY})\\ &\geqslant D(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q))\\ &= -H(p)-p\log q-(1-p)\log(1-q)\\ &> -1-\delta_P\log q,\end{aligned}$$

其中第二个不等号依据数据处理不等式。因此一个符合条件的 δ_Q 取值为

$$\delta_Q \triangleq \exp\left(-\frac{1}{\delta_P}[D(P_{XY}\|P_XP_Y)+1]\right), \tag{A-50}$$

且其满足 $0 \leqslant \delta_Q < q = Q_{XY}(x,y) < 1$. 设 $Q_{XY} = Q[f,g,\alpha,\beta]$，则 $f^{\mathsf{T}}(x)g(y)-\alpha(x)-\beta(y)$ 有界，即存在 $M_1 > 0$ 使得 $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}$,

$$\left|f^{\mathsf{T}}(x)g(y)-\alpha(x)-\beta(y)\right| \leqslant M_1. \tag{A-51}$$

注意到 Q_{XY} 参数化的方式不唯一，例如恒等式 $Q[f,g,\alpha,\beta] = Q[f+c,g,\alpha+c^{\mathsf{T}}g,\beta]$ 对任意常向量 c 成立。故可因此可不妨假设 $\mathbb{E}_{P_X}[f(X)] = \mathbb{E}_{P_Y}[g(Y)] = 0$ 及 $\mathbb{E}_{P_X}[\alpha(X)] = 0$。由 Jensen 不等式可知，对任意 $(x,y) \in \mathcal{X}\times\mathcal{Y}$ 有

$$|\beta(y)| = \left|\mathbb{E}_{P_X}\left[f^{\mathsf{T}}(X)g(y)-\alpha(X)-\beta(y)\right]\right| \leqslant \mathbb{E}_{P_X}\left[\left|f^{\mathsf{T}}(X)g(y)-\alpha(X)-\beta(y)\right|\right] \leqslant M_1,$$

以及

$$\begin{aligned}
|\alpha(x)| &\leqslant \left|\alpha(x)+\mathbb{E}_{P_Y}[\beta(Y)]\right| + \left|\mathbb{E}_{P_Y}[\beta(Y)]\right| \\
&\leqslant \left|\mathbb{E}_{P_Y}\left[f^{\mathsf{T}}(x)g(Y)-\alpha(x)-\beta(Y)\right]\right| + M_1 \\
&\leqslant \mathbb{E}_{P_Y}\left[\left|f^{\mathsf{T}}(x)g(Y)-\alpha(x)-\beta(Y)\right|\right] + M_1 \\
&\leqslant 2M_1.
\end{aligned}$$

由此可得

$$\left|f^{\mathsf{T}}(x)g(y)\right| \leqslant \left|f^{\mathsf{T}}(x)g(y)-\alpha(x)-\beta(y)\right| + |\alpha(x)| + |\beta(y)| \leqslant 4M_1,$$

即 $\mathbf{F}\mathbf{G}^{\mathsf{T}}$ 中所有元素有上界 $4M_1$，从而由范数等价性[80] 知

$$\left\|\mathbf{F}\mathbf{G}^{\mathsf{T}}\right\|_2 \leqslant \sqrt{|\mathcal{X}||\mathcal{Y}|}\left\|\mathbf{F}\mathbf{G}^{\mathsf{T}}\right\|_{\max} \leqslant 4\sqrt{|\mathcal{X}||\mathcal{Y}|}M_1. \tag{A-52}$$

设 $\mathbf{F}\mathbf{G}^{\mathsf{T}}$ 有精简奇异值分解 (Compact Singular Value Decomposition) $\mathbf{F}\mathbf{G}^{\mathsf{T}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$，其中 $\mathbf{\Sigma}$ 为由所有 $\mathbf{F}\mathbf{G}^{\mathsf{T}}$ 的非零奇异值构成的对角阵，则可构造 $\hat{\mathbf{F}} = \mathbf{U}\mathbf{\Sigma}^{1/2}$ 以及 $\hat{\mathbf{G}} = \mathbf{V}\mathbf{\Sigma}^{1/2}$，使得

$$\hat{\mathbf{F}}\hat{\mathbf{G}}^{\mathsf{T}} = \mathbf{F}\mathbf{G}^{\mathsf{T}}, \left\|\hat{\mathbf{F}}\right\|_2 = \left\|\hat{\mathbf{G}}\right\|_2 = \left\|\mathbf{F}\mathbf{G}^{\mathsf{T}}\right\|_2^{1/2}. \tag{A-53}$$

令 $\hat{f},\hat{g}$ 分别为 $\hat{\mathbf{F}},\hat{\mathbf{G}}$ 对应的函数，则有 $Q[f,g,\alpha,\beta] = Q[\hat{f},\hat{g},\alpha,\beta]$。由于 $\|\hat{\mathbf{F}}\|_2$ 与 $\|\hat{\mathbf{G}}\|_2$ 均有界，由范数等价性知 $\|\hat{\mathbf{F}}\|_{\mathsf{F}} \leqslant M_2, \|\hat{\mathbf{G}}\|_{\mathsf{F}} \leqslant M_2$，其中 $M_2 \triangleq 2\sqrt{|\mathcal{X}||\mathcal{Y}|M_1}$。

故 $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}$ 有

$$\|\hat{f}(x)\| \leqslant \|\hat{\mathbf{F}}\|_{\mathsf{F}} \leqslant M_2, \|\hat{g}(y)\| \leqslant \|\hat{\mathbf{G}}\|_{\mathsf{F}} \leqslant M_2. \tag{A-54}$$

令 $M \triangleq \max\{2M_1, M_2\}$，则 M 仅取决于 P_{XY}，从而可得 $Q_{XY} = Q[\hat{f}, \hat{g}, \alpha, \beta]$，其中

$$\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{\|\hat{f}(x)\|, \|\hat{g}(y)\|, |\alpha(x)|, |\beta(y)|\right\} \leqslant M(P_{XY}). \qquad \square$$

A.8　引理 3.4 的证明

证明 设 $R_{XY} \in \mathcal{E}_k$，则存在参数 f, g, α, β 使得 $R_{XY} = Q[f, g, \alpha, \beta]$。故对任意 $(x,y) \in \mathcal{X} \times \mathcal{Y}$，有，

$$\log \frac{R_{XY}(x,y)}{R_X(x)R_Y(y)} = f^{\mathsf{T}}(x)g(y) - \alpha(x) - \beta(y). \tag{A-55}$$

其等价于

$$\mathbf{\Gamma} = \mathbf{F}\mathbf{G}^{\mathsf{T}} - \mathbf{1}_{|\mathcal{X}|}\boldsymbol{\beta}^{\mathsf{T}} - \boldsymbol{\alpha}\mathbf{1}_{|\mathcal{Y}|}^{\mathsf{T}}, \tag{A-56}$$

其中 $\mathbf{1}_m$ 为所有元素均为 1 的 m 维向量。若 $\operatorname{rank}(\mathbf{\Gamma}) \leqslant k$，可找到 $(\mathbf{F}, \mathbf{G})$ 对使 $\mathbf{\Gamma} = \mathbf{F}\mathbf{G}^{\mathsf{T}}$ 以及 $R_{XY} = Q[f, g, 0, 0]$ 成立。此外，由 (A-56) 可得

$$\operatorname{rank}(\mathbf{\Gamma}) \leqslant \operatorname{rank}(\mathbf{F}\mathbf{G}^{\mathsf{T}}) + 2 \leqslant k + 2. \tag{A-57}$$

故当 $\operatorname{rank}(\mathbf{\Gamma}) > k + 2$ 时有 $R_{XY} \notin \mathcal{E}_k$。 $\square$

附录 B　第 4 章中的证明

B.1　引理 4.1 的证明

设 $\lambda_1, \cdots, \lambda_k$ 共有 q 个不同的取值，定义下标 $i_0, \cdots, i_q$ 使其满足 $0 = i_0 < \cdots < i_q = k$ 及

$$\lambda_{i_0+1} = \lambda_{i_1} > \lambda_{i_1+1} = \lambda_{i_2} > \cdots > \lambda_{i_{q-1}+1} = \lambda_{i_q} > \lambda_{i_q+1}.$$

由此可得

$$\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_k(\tau)\right\} = \sum_{j=1}^{k}\mathbf{v}_j^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{v}_j(\tau) = \sum_{s=1}^{q}\sum_{i=i_{s-1}+1}^{i_s}\mathbf{v}_i^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{v}_i(\tau), \tag{B-1}$$

其中 $\mathbf{v}_j(\tau)$ 表示 $\mathbf{V}_k$ 的第 j 列。接下来，我们考虑上述求和式中 $s = 1$ 的项，即

$$\sum_{i=i_0+1}^{i_1}\mathbf{v}_i^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{v}_i(\tau) = \operatorname{tr}\left\{\mathbf{V}_{i_1}^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_{i_1}(\tau)\right\},$$

其中 $\mathbf{V}_{i_1}(\tau) \in \mathbb{R}^{d\times i_1}$ 由矩阵 $\mathbf{V}_k(\tau)$ 的前 i_1 列构成。注意到由于 $\mathbf{A}(\tau)$ 为解析函数，因此存在对称阵 $\mathbf{A}''$ 使得

$$\mathbf{A}(\tau) = \mathbf{A} + \tau\mathbf{A}' + \tau^2\mathbf{A}'' + o\left(\tau^2\right).$$

此外，由 $\mathbf{A}(\tau)$ 的解析性可知对应特征空间 $\mathbf{V}_k(\tau)$ 的解析性[94]，从而我们可将 $\mathbf{V}_{i_1}(\tau)$ 展开为

$$\mathbf{V}_{i_1}(\tau) = \hat{\mathbf{V}}_{i_1} + \tau\mathbf{V}'_{i_1} + \tau^2\mathbf{V}''_{i_1} + o\left(\tau^2\right),$$

其中 $\hat{\mathbf{V}}_{i_1}$、$\mathbf{V}'_{i_1}$ 及 $\mathbf{V}''_{i_1}$ 均为 $\mathbb{R}^{d\times i_1}$ 中的矩阵。进一步地, 由于 $\hat{\mathbf{V}}_{i_1}$ 的各列构成矩阵 $\mathbf{A}$ 与特征值 λ_1 对应的特征子空间的一组标准正交基，我们有 $\mathbf{A}\hat{\mathbf{V}}_{i_1} = \lambda_1\hat{\mathbf{V}}_{i_1}$。从而根据 $\mathbf{V}_{i_1}^{\mathsf{T}}(\tau)\mathbf{V}_{i_1}(\tau) = \mathbf{I}_{i_1}$ 可推知

$$\begin{aligned}\mathbf{I}_{i_1} &= \mathbf{V}_{i_1}^{\mathsf{T}}(\tau)\mathbf{V}_{i_1}(\tau)\\ &= \hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\hat{\mathbf{V}}_{i_1} + \tau\left(\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}'_{i_1} + \mathbf{V}_{i_1}'^{\mathsf{T}}\hat{\mathbf{V}}_{i_1}\right) + \tau^2\left(\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}''_{i_1} + \mathbf{V}_{i_1}''^{\mathsf{T}}\hat{\mathbf{V}}_{i_1} + \mathbf{V}_{i_1}'^{\mathsf{T}}\mathbf{V}'_{i_1}\right) + o(\tau^2),\end{aligned}$$

因此可得

$$\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}'_{i_1} + \mathbf{V}_{i_1}'^{\mathsf{T}}\hat{\mathbf{V}}_{i_1} = \mathbf{O}_{i_1}, \tag{B-2a}$$

$$\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}_{i_1}'' + \mathbf{V}_{i_1}''^{\mathsf{T}}\hat{\mathbf{V}}_{i_1} + \mathbf{V}_{i_1}'^{\mathsf{T}}\mathbf{V}_{i_1}' = \mathbf{O}_{i_1}, \tag{B-2b}$$

其中 $\mathbf{I}_{i_1}$ 及 $\mathbf{O}_{i_1}$ 分别表示 $\mathbb{R}^{i_1\times i_1}$ 中的单位阵及零矩阵。因此，我们有

$$\begin{aligned}
&\mathbf{V}_{i_1}^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_{i_1}(\tau)\\
&\quad= \left(\hat{\mathbf{V}}_{i_1} + \tau\mathbf{V}_{i_1}' + \tau^2\mathbf{V}_{i_1}''\right)^{\mathsf{T}}\mathbf{A}\left(\hat{\mathbf{V}}_{i_1} + \tau\mathbf{V}_{i_1}' + \tau^2\mathbf{V}_{i_1}''\right) + o(\tau^2)\\
&\quad= \hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}\hat{\mathbf{V}}_{i_1} + \tau\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}\mathbf{V}_{i_1}' + \mathbf{V}_{i_1}'^{\mathsf{T}}\mathbf{A}\hat{\mathbf{V}}_{i_1} + \tau^2\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}\mathbf{V}_{i_1}'' + \mathbf{V}_{i_1}'^{\mathsf{T}}\mathbf{A}\mathbf{V}_{i_1}' + \mathbf{V}_{i_1}''^{\mathsf{T}}\mathbf{A}\hat{\mathbf{V}}_{i_1} + o(\tau^2)\\
&\quad= \lambda_1\cdot\mathbf{I}_{i_1} + \lambda_1\cdot\tau\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}_{i_1}' + \mathbf{V}_{i_1}'^{\mathsf{T}}\hat{\mathbf{V}}_{i_1} + \tau^2\lambda_1\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}_{i_1}'' + \mathbf{V}_{i_1}'^{\mathsf{T}}\mathbf{A}\mathbf{V}_{i_1}' + \lambda_1\mathbf{V}_{i_1}''^{\mathsf{T}}\hat{\mathbf{V}}_{i_1} + o(\tau^2)\\
&\quad= \lambda_1\cdot\mathbf{I}_{i_1} - \tau^2\mathbf{V}_{i_1}'^{\mathsf{T}}(\lambda_1\mathbf{I}_d - \mathbf{A})\mathbf{V}_{i_1}' + o(\tau^2),
\end{aligned} \tag{B-3}$$

其中倒数第二个等号根据 $\mathbf{A}\hat{\mathbf{V}}_{i_1} = \lambda_1\hat{\mathbf{V}}_{i_1}$ 推得，最后一个等号基于 (B-2) 得出，式中 $\mathbf{I}_d$ 表示 $\mathbb{R}^{d\times d}$ 中的单位阵。

进一步，我们定义矩阵

$$\mathbf{\Lambda}_{i_1}(\tau) \triangleq \mathrm{diag}\{\lambda_1(\tau), \cdots, \lambda_{i_1}(\tau)\},$$

其中 $\lambda_1(\tau), \cdots, \lambda_{i_1}(\tau)$ 为矩阵 $\mathbf{A}(\tau)$ 的前 i_1 个特征值。根据 $\mathbf{A}(\tau)$ 的解析性可得 $\mathbf{\Lambda}_{i_1}(\tau)$ 的解析性，因此其可展开成为

$$\mathbf{\Lambda}_{i_1}(\tau) = \lambda_1\mathbf{I}_{i_1} + \tau\mathbf{\Lambda}_{i_1}' + \tau^2\mathbf{\Lambda}_{i_1}'' + o(\tau^2),$$

其中 $\mathbf{\Lambda}_{i_1}'$ 及 $\mathbf{\Lambda}_{i_1}''$ 均为对角阵。于是根据 $\mathbf{A}(\tau)\mathbf{V}_{i_1}(\tau) = \mathbf{V}_{i_1}(\tau)\mathbf{\Lambda}_{i_1}(\tau)$ 可知

$$\begin{aligned}
&\left(\mathbf{A} + \tau\mathbf{A}' + \tau^2\mathbf{A}''\right)\left(\hat{\mathbf{V}}_{i_1} + \tau\mathbf{V}_{i_1}' + \tau^2\mathbf{V}_{i_1}''\right)\\
&\qquad= \left(\hat{\mathbf{V}}_{i_1} + \tau\mathbf{V}_{i_1}' + \tau^2\mathbf{V}_{i_1}''\right)\left(\lambda_1\mathbf{I}_{i_1} + \tau\mathbf{\Lambda}_{i_1}' + \tau^2\mathbf{\Lambda}_{i_1}''\right) + o(\tau^2).
\end{aligned}$$

比较等式两边 τ 的一阶项，可得

$$\mathbf{A}'\hat{\mathbf{V}}_{i_1} + \mathbf{A}\mathbf{V}_{i_1}' = \lambda_1\mathbf{V}_{i_1}' + \hat{\mathbf{V}}_{i_1}\mathbf{\Lambda}_{i_1}',$$

由此推出

$$(\lambda_1\mathbf{I}_d - \mathbf{A})\mathbf{V}_{i_1}' = \mathbf{A}'\hat{\mathbf{V}}_{i_1} - \hat{\mathbf{V}}_{i_1}\mathbf{\Lambda}_{i_1}'. \tag{B-4}$$

在上式等号两边同时左乘 $\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}$，并利用等式 $\mathbf{A}\hat{\mathbf{V}}_{i_1} = \lambda_1\hat{\mathbf{V}}_{i_1}$，化简可得

$$\mathbf{\Lambda}_{i_1}' = \hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}'\hat{\mathbf{V}}_{i_1}. \tag{B-5}$$

于是我们可将 $\left[\mathbf{V}_{i_1}'^{\mathsf{T}}(\lambda_1\mathbf{I}_d-\mathbf{A})\mathbf{V}_{i_1}'\right]$ of (B-3) 表达为

$$\begin{aligned}
\mathbf{V}_{i_1}'^{\mathsf{T}}(\lambda_1\mathbf{I}_d-\mathbf{A})\mathbf{V}_{i_1}' &= \left[(\lambda_1\mathbf{I}_d-\mathbf{A})\mathbf{V}_{i_1}'\right]^{\mathsf{T}}\mathbf{V}_{i_1}' \\
&= \left(\mathbf{A}'\hat{\mathbf{V}}_{i_1}-\hat{\mathbf{V}}_{i_1}\mathbf{\Lambda}_{i_1}'\right)^{\mathsf{T}}\mathbf{V}_{i_1}' && \text{(B-6a)} \\
&= \left(\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}'-\mathbf{\Lambda}_{i_1}'\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\right)\mathbf{V}_{i_1}' && \text{(B-6b)} \\
&= \hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}'\left[\left(\mathbf{I}_d-\hat{\mathbf{V}}_{i_1}\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\right)+\hat{\mathbf{V}}_{i_1}\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\right]\mathbf{V}_{i_1}'-\mathbf{\Lambda}_{i_1}'\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}_{i_1}' \\
&= \hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}'\left(\mathbf{I}_d-\hat{\mathbf{V}}_{i_1}\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\right)\mathbf{V}_{i_1}'+\left(\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}'\hat{\mathbf{V}}_{i_1}-\mathbf{\Lambda}_{i_1}'\right)\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{V}_{i_1}' \\
&= \hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\mathbf{A}'\left(\mathbf{I}_d-\hat{\mathbf{V}}_{i_1}\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\right)\mathbf{V}_{i_1}', && \text{(B-6c)}
\end{aligned}$$

其中 (B-6a) 基于 (B-4)，(B-6c) 基于 (B-5)。进一步，由矩阵 $\mathbf{A}$ 的特征值分解有

$$\mathbf{A}=\lambda_1\sum_{j=1}^{i_1}\hat{\mathbf{v}}_j\hat{\mathbf{v}}_j^{\mathsf{T}}+\sum_{j=i_1+1}^{d}\lambda_j\mathbf{v}_j\mathbf{v}_j^{\mathsf{T}},$$

其中 $\hat{\mathbf{v}}_j$ 为矩阵 $\hat{\mathbf{V}}_{i_1}$ 的第 j 列 $(1\leqslant j\leqslant i_1)$。类似地，我们有

$$\mathbf{I}_d=\sum_{j=1}^{i_1}\hat{\mathbf{v}}_j\hat{\mathbf{v}}_j^{\mathsf{T}}+\sum_{j=i_1+1}^{d}\mathbf{v}_j\mathbf{v}_j^{\mathsf{T}}.$$

由此可得

$$\lambda_1\mathbf{I}_d-\mathbf{A}=\sum_{j=i_1+1}^{d}(\lambda_1-\lambda_j)\mathbf{v}_j\mathbf{v}_j^{\mathsf{T}},$$

及其 Moore–Penrose 伪逆

$$(\lambda_1\mathbf{I}_d-\mathbf{A})^{\dagger}=\sum_{j=i_1+1}^{d}\frac{\mathbf{v}_j\mathbf{v}_j^{\mathsf{T}}}{\lambda_1-\lambda_j}. \tag{B-7}$$

于是可推出

$$\mathbf{I}_d-\hat{\mathbf{V}}_{i_1}\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}=\sum_{j=i_1+1}^{d}\mathbf{v}_j\mathbf{v}_j^{\mathsf{T}}=(\lambda_1\mathbf{I}_d-\mathbf{A})^{\dagger}(\lambda_1\mathbf{I}_d-\mathbf{A}),$$

从而有

$$\begin{aligned}
\left(\mathbf{I}_d-\hat{\mathbf{V}}_{i_1}\hat{\mathbf{V}}_{i_1}^{\mathsf{T}}\right)\mathbf{V}_{i_1}' &= (\lambda_1\mathbf{I}_d-\mathbf{A})^{\dagger}(\lambda_1\mathbf{I}_d-\mathbf{A})\mathbf{V}_{i_1}' && \text{(B-8a)} \\
&= (\lambda_1\mathbf{I}_d-\mathbf{A})^{\dagger}\mathbf{A}'\hat{\mathbf{V}}_{i_1}-(\lambda_1\mathbf{I}_d-\mathbf{A})^{\dagger}\hat{\mathbf{V}}_{i_1}\mathbf{\Lambda}_{i_1}' && \text{(B-8b)}
\end{aligned}$$

$$= (\lambda_1 \mathbf{I}_d - \mathbf{A})^{\dagger} \mathbf{A}' \hat{\mathbf{V}}_{i_1}, \tag{B-8c}$$

其中 (B-8b) 基于 (B-4)。为导出 (B-8c)，注意到由于对所有 $i \leqslant i_1 < j$，$\hat{\mathbf{v}}_i$ 与 $\mathbf{v}_j$ 正交，因此 $(\lambda_1 \mathbf{I}_d - \mathbf{A})^{\dagger} \hat{\mathbf{V}}_{i_1}$ 为零矩阵 [参见 (B-7)]。

综合 (B-3)、(B-6) 及 (B-8) 的结果，可得

$$\mathbf{V}_{i_1}^{\mathsf{T}}(\tau) \mathbf{A} \mathbf{V}_{i_1}(\tau) = \lambda_1 \cdot \mathbf{I}_{i_1} - \tau^2 \hat{\mathbf{V}}_{i_1}^{\mathsf{T}} \mathbf{A}' (\lambda_1 \mathbf{I}_d - \mathbf{A})^{\dagger} \mathbf{A}' \hat{\mathbf{V}}_{i_1} + o(\tau^2), \tag{B-9}$$

由此推出

$$\begin{aligned}
\operatorname{tr}\left\{ \mathbf{V}_{i_1}^{\mathsf{T}}(\tau) \mathbf{A} \mathbf{V}_{i_1}(\tau) \right\} &= \lambda_1 \cdot i_1 - \tau^2 \operatorname{tr}\left\{ \hat{\mathbf{V}}_{i_1}^{\mathsf{T}} \mathbf{A}' (\lambda_1 \mathbf{I}_d - \mathbf{A})^{\dagger} \mathbf{A}' \hat{\mathbf{V}}_{i_1} \right\} + o(\tau^2) \\
&= \lambda_1 \cdot i_1 - \tau^2 \sum_{i=1}^{i_1} \hat{\mathbf{v}}_i^{\mathsf{T}} \mathbf{A}' (\lambda_1 \mathbf{I}_d - \mathbf{A})^{\dagger} \mathbf{A}' \hat{\mathbf{v}}_i + o(\tau^2) \\
&= \lambda_1 \cdot i_1 - \tau^2 \sum_{i=1}^{i_1} \sum_{j=i_1+1}^{d} \frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j \right)^2}{\lambda_1 - \lambda_j} + o(\tau^2) && \text{(B-10a)} \\
&= \lambda_1 \cdot i_1 - \tau^2 \sum_{j=i_1+1}^{d} \frac{\left\| \hat{\mathbf{V}}_{i_1}^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j \right\|^2}{\lambda_1 - \lambda_j} + o(\tau^2) && \text{(B-10b)} \\
&= \lambda_1 \cdot i_1 - \tau^2 \sum_{j=i_1+1}^{d} \frac{\left\| \mathbf{V}_{i_1}^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j \right\|^2}{\lambda_1 - \lambda_j} + o(\tau^2) && \text{(B-10c)} \\
&= \lambda_1 \cdot i_1 - \tau^2 \sum_{i=1}^{i_1} \sum_{j=i_1+1}^{d} \frac{\left(\mathbf{v}_i^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j \right)^2}{\lambda_1 - \lambda_j} + o(\tau^2), && \text{(B-10d)}
\end{aligned}$$

其中 (B-10a) 依据 (B-7)，等式 (B-10c) 中的 $\mathbf{V}_{i_1}$ 定义为 $\mathbf{V}_{i_1} \triangleq [\mathbf{v}_1, \cdots, \mathbf{v}_{i_1}] \in \mathbb{R}^{d \times i_1}$。此外，等式 (B-10c) 成立是由于，矩阵 $\mathbf{V}_{i_1}$ 及 $\hat{\mathbf{V}}_{i_1}$ 的各列均构成 $\mathbf{A}$ 与特征值 λ_1 对应的特征子空间的一组正交基。

基于类似的推导，对任意的 s 我们有

$$\sum_{i=i_{s-1}+1}^{i_s} \mathbf{v}_i^{\mathsf{T}}(\tau) \mathbf{A} \mathbf{v}_i(\tau) = \lambda_{i_s}(i_s - i_{s-1}) - \tau^2 \sum_{i=i_{s-1}+1}^{i_s} \sum_{j:\lambda_j \neq \lambda_i} \frac{\left(\mathbf{v}_i^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j \right)^2}{\lambda_i - \lambda_j} + o(\tau^2). \tag{B-11}$$

因此，结合 (B-1) 及 (B-11) 可得

$$\operatorname{tr}\left\{ \mathbf{V}_k^{\mathsf{T}}(\tau) \mathbf{A} \mathbf{V}_k(\tau) \right\} = \sum_{s=1}^{q} \sum_{i=i_{s-1}+1}^{i_s} \mathbf{v}_i^{\mathsf{T}}(\tau) \mathbf{A} \mathbf{v}_i(\tau)$$

$$
\begin{aligned}
&= \sum_{s=1}^{q} \lambda_{i_s}(i_s - i_{s-1}) - \tau^2 \sum_{i=1}^{k} \sum_{j=k+1}^{d} \frac{\left(\mathbf{v}_i^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j\right)^2}{\lambda_i - \lambda_j} + o(\tau^2) \\
&= \operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}} \mathbf{A} \mathbf{V}_k\right\} - \tau^2 \sum_{i=1}^{k} \sum_{j=k+1}^{d} \frac{\left(\mathbf{v}_i^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j\right)^2}{\lambda_i - \lambda_j} + o(\tau^2),
\end{aligned}
$$

即欲证结论。

B.2　引理 4.2 的证明

设 $\lambda_1, \cdots, \lambda_k$ 有 q 个不同的取值，定义下标 $i_0, \cdots, i_q$ 使其满足 $0 = i_0 < \cdots < i_{q-1} < k < k+1 \leqslant i_q$ 及

$$
\lambda_{i_{s-1}+1} = \lambda_{i_s} > \lambda_{i_s+1}, \quad 0 \leqslant s \leqslant q.
$$

首先考虑 $q = 1$ 的情形，从而有 $k < i_1$。由 (B-9) 可知

$$
\mathbf{V}_k^{\mathsf{T}}(\tau) \mathbf{A} \mathbf{V}_k(\tau) = \lambda_1 \cdot \mathbf{I}_k - \tau^2 \hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{A}' (\lambda_1 \mathbf{I}_d - \mathbf{A})^{\dagger} \mathbf{A}' \hat{\mathbf{V}}_k + o(\tau^2),
$$

从而 [参见 (B-10)]

$$
\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}} \mathbf{A} \mathbf{V}_k\right\} = k\lambda_1 - \tau^2 \sum_{i=1}^{k} \sum_{j \in \mathcal{I}_k^{\mathsf{c}}} \frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}} \mathbf{A}' \mathbf{v}_j\right)^2}{\lambda_1 - \lambda_j} + o(\tau^2)
$$

为了求解 $\hat{\mathbf{v}}_i$ $(1 \leqslant i \leqslant k)$，注意到

$$
\begin{aligned}
&\mathbf{V}_k^{\mathsf{T}}(\tau) \mathbf{A}(\tau) \mathbf{V}_k(\tau) \\
&\quad = \left(\hat{\mathbf{V}}_k + \tau \mathbf{V}_k' + \tau^2 \mathbf{V}_k''\right)^{\mathsf{T}} \left(\mathbf{A} + \tau \mathbf{A}' + \tau^2 \mathbf{A}''\right) \left(\hat{\mathbf{V}}_k + \tau \mathbf{V}_k' + \tau^2 \mathbf{V}_k''\right) + o(\tau^2) \\
&\quad = \hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{A} \hat{\mathbf{V}}_k + \tau \left(\hat{\mathbf{V}}_k \mathbf{A} \mathbf{V}_k' + \mathbf{V}_k' \mathbf{A} \hat{\mathbf{V}}_k + \hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{A} \hat{\mathbf{V}}_k\right) + o(\tau) \\
&\quad = \lambda_1 \mathbf{I}_k + \tau \left[\lambda_1 \left(\hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{V}_k' + \mathbf{V}_k'^{\mathsf{T}} \hat{\mathbf{V}}_k\right) + \hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{A}' \hat{\mathbf{V}}_k\right] + o(\tau) \\
&\quad = \lambda_1 \mathbf{I}_k + \tau \hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{A}' \hat{\mathbf{V}}_k + o(\tau),
\end{aligned}
$$

其中第三个等式的推导基于 $\mathbf{A}\hat{\mathbf{V}}_k = \lambda_1 \hat{\mathbf{V}}_k$。为导出最后一个等式，注意到根据 (B-2a)，$\left(\hat{\mathbf{V}}_k^{\mathsf{T}} \mathbf{V}_k' + \mathbf{V}_k'^{\mathsf{T}} \hat{\mathbf{V}}_k\right)$ 为零矩阵。

由于矩阵 $\hat{\mathbf{V}}_k$ 各列为矩阵 $\mathbf{A}$ 对应于特征值 λ_1 的特征子空间中 k 个单位正交的向量，我们可以将 $\hat{\mathbf{V}}_k$ 表达为 $\hat{\mathbf{V}}_k = \mathbf{V}_{i_1} \mathbf{U}$ 的形式，其中 $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_k] \in \mathbb{R}^{i_1 \times k}$ 满

足 $\mathbf{U}^{\mathsf{T}}\mathbf{U}=\mathbf{I}_k$。此外，由特征向量的定义，$\mathbf{V}_k(\tau)$ 为所有各列单位正交的 $d\times k$ 矩阵中，使得

$$\begin{aligned}\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}(\tau)\mathbf{V}_k(\tau)\right\}&=k\lambda_1+\tau\operatorname{tr}\left\{\hat{\mathbf{V}}_k^{\mathsf{T}}\mathbf{A}'\hat{\mathbf{V}}_k\right\}+o(\tau)\\&=k\lambda_1+\tau\operatorname{tr}\left\{\mathbf{U}^{\mathsf{T}}\mathbf{V}_{i_1}^{\mathsf{T}}\mathbf{A}'\mathbf{V}_{i_1}\mathbf{U}\right\}+o(\tau)\end{aligned}$$

最大化的矩阵。

因此，$\mathbf{U}$ 可表示为如下优化问题的最优解：

$$\begin{aligned}\underset{\mathbf{U}'}{\text{maximize}}\quad&\operatorname{tr}\left\{\mathbf{U}'^{\mathsf{T}}\mathbf{V}_{i_1}^{\mathsf{T}}\mathbf{A}'\mathbf{V}_{i_1}\mathbf{U}'\right\}\\\text{subject to}\quad&\mathbf{U}'\in\mathbb{R}^{i_1\times k},\quad\mathbf{U}'^{\mathsf{T}}\mathbf{U}'=\mathbf{I}_k,\end{aligned}$$

从而可知 $\mathbf{u}_1,\cdots,\mathbf{u}_k$ 为矩阵 $\mathbf{V}_{i_1}^{\mathsf{T}}\mathbf{A}'\mathbf{V}_{i_1}$ 的前 k 个特征向量。

由此可得

$$\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}\mathbf{A}\mathbf{V}_k\right\}=k\lambda_1-\tau^2\sum_{i=1}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_1-\lambda_j}+o(\tau^2)\tag{B-12}$$

其中 $\hat{\mathbf{v}}_i=\mathbf{V}_{i_1}\mathbf{u}_i$ $(1\leqslant i\leqslant k)$，且 $\mathbf{u}_1,\cdots,\mathbf{u}_k$ 为矩阵 $\mathbf{V}_{i_1}^{\mathsf{T}}\mathbf{A}'\mathbf{V}_{i_1}$ 的前 k 个特征向量。

类似地，对 $q>1$ 我们有

$$\sum_{i=l}^{k}\mathbf{v}_i^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{v}_i(\tau)=(k-l+1)\lambda_k-\tau^2\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_k-\lambda_j}+o(\tau^2),$$

其中 l 为 $\mathcal{I}_k$ 的最小元，$\hat{\mathbf{v}}_i$ 定义为

$$\hat{\mathbf{v}}_i\triangleq\mathbf{V}_{\mathcal{I}_k}\mathbf{u}_{i-l+1},\quad l\leqslant i\leqslant k,$$

其中 $\mathbf{u}_1,\cdots,\mathbf{u}_{k-l+1}$ 为 $\mathbf{V}_{\mathcal{I}_k}^{\mathsf{T}}\mathbf{A}'\mathbf{V}_{\mathcal{I}_k}$ 的前 $k-l+1$ 个特征向量。由此可得

$$\begin{aligned}&\operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_k(\tau)\right\}\\&\quad=\operatorname{tr}\left\{\mathbf{V}_{l-1}^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{V}_{l-1}(\tau)\right\}+\sum_{i=l}^{k}\mathbf{v}_i^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{v}_i(\tau)\\&\quad=\operatorname{tr}\left\{\mathbf{V}_{l-1}^{\mathsf{T}}\mathbf{A}\mathbf{V}_{l-1}\right\}-\tau^2\sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\left(\mathbf{v}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_i-\lambda_j}+\sum_{i=l}^{k}\mathbf{v}_i^{\mathsf{T}}(\tau)\mathbf{A}\mathbf{v}_i(\tau)+o(\tau^2)\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{tr}\left\{\mathbf{V}_{l-1}^{\mathsf{T}}\mathbf{A}\mathbf{V}_{l-1}\right\} - \tau^2 \sum_{i=1}^{l-1}\sum_{j=l}^{d} \frac{\left(\mathbf{v}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_i - \lambda_j} \\
&\quad + (k-l+1)\lambda_r - \tau^2 \sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}} \frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_k - \lambda_j} + o(\tau^2) \\
&= \operatorname{tr}\left\{\mathbf{V}_k^{\mathsf{T}}\mathbf{A}\mathbf{V}_k\right\} - \tau^2 \sum_{i=1}^{l-1}\sum_{j=l}^{d} \frac{\left(\mathbf{v}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_i - \lambda_j} - \tau^2 \sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}} \frac{\left(\hat{\mathbf{v}}_i^{\mathsf{T}}\mathbf{A}'\mathbf{v}_j\right)^2}{\lambda_k - \lambda_j} + o(\tau^2),
\end{aligned}
$$

其中第二个等号可由引理 4.1推出。

B.3　引理 4.3 的证明

首先定义 $p_{\min} \triangleq \min\{P_{XY}(x,y)\colon (x,y)\in\mathcal{X}\times\mathcal{Y}, P_{XY}(x,y)>0\}$，则对任意 $\hat{P}_{XY}\in\mathcal{N}(\epsilon)$，根据 Pinsker 不等式[9] 我们有

$$
\begin{aligned}
\sum_{x\in\mathcal{X},y\in\mathcal{Y}} [\Gamma(y,x)]^2 &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}\colon P_{XY}(x,y)>0} \frac{\left[\hat{P}_{XY}(x,y) - P_{XY}(x,y)\right]^2}{\epsilon P_{XY}(x,y)} && \text{(B-13a)} \\
&\leqslant \frac{1}{\epsilon p_{\min}} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \left[\hat{P}_{XY}(x,y) - P_{XY}(x,y)\right]^2 && \text{(B-13b)} \\
&\leqslant \frac{1}{\epsilon p_{\min}} \left(\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \left|\hat{P}_{XY}(x,y) - P_{XY}(x,y)\right| \right)^2 && \text{(B-13c)} \\
&\leqslant \frac{2D(\hat{P}_{XY}\|P_{XY})}{\epsilon p_{\min}} && \text{(B-13d)} \\
&\leqslant \frac{2}{p_{\min}\alpha_k}. && \text{(B-13e)}
\end{aligned}
$$

故对所有 $(x,y)\in\mathcal{X}\times\mathcal{Y}$ 有

$$
|\Gamma(y,x)| < \sqrt{\frac{2}{p_{\min}\alpha_k}}. \tag{B-14}
$$

因此，由 (4-16) 可知

$$
\begin{aligned}
|\Xi(y,x)| &\leqslant \frac{1}{p_{\min}}|\Gamma(y,x)| + \frac{1}{p_{\min}}\left| \frac{1}{P_X(x)} \sum_{y'\in\mathcal{Y}} \sqrt{P_{XY}(x,y')}\Gamma(y',x) \right. && \text{(B-15)} \\
&\quad \left. + \frac{1}{P_Y(y)} \sum_{x'\in\mathcal{X}} \sqrt{P_{XY}(x',y)}\Gamma(y,x') \right| && \text{(B-16)}
\end{aligned}
$$

$$\leqslant \frac{1}{p_{\min}}|\Gamma(y,x)| + \frac{|\mathcal{X}|+|\mathcal{Y}|}{p_{\min}^2}\cdot \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|\Gamma(y,x)| \tag{B-17}$$

$$\leqslant \frac{1+|\mathcal{X}|+|\mathcal{Y}|}{p_{\min}^2}\cdot \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|\Gamma(y,x)| \tag{B-18}$$

$$\leqslant \frac{1+|\mathcal{X}|+|\mathcal{Y}|}{p_{\min}^3}\sqrt{\frac{2}{\alpha_k}}, \tag{B-19}$$

其中最后的不等式基于 (B-14) 及 $p_{\min} \leqslant 1$。于是我们有 $\|\boldsymbol{\Xi}\|_{\mathrm{F}} \leqslant C$，其中 $C \triangleq \frac{1+|\mathcal{X}|+|\mathcal{Y}|}{p_{\min}^3}\sqrt{\frac{2|\mathcal{X}||\mathcal{Y}|}{\alpha_k}}$ 仅仅依赖于 P_{XY}。

欲证 (4-17)，我们首先引入记号 $\tau \triangleq \sqrt{\epsilon}$ 以便表达。根据 (4-15)，经验边缘分布与真实边缘分布的差可表示为

$$\hat{P}_X(x) - P_X(x) = \tau\sqrt{P_X(x)}\Gamma_X(x), \tag{B-20}$$

$$\hat{P}_Y(y) - P_Y(y) = \tau\sqrt{P_Y(y)}\Gamma_Y(y), \tag{B-21}$$

其中

$$\Gamma_X(x) \triangleq \sum_{y\in\mathcal{Y}}\sqrt{P_{Y|X}(y|x)}\Gamma(y,x), \tag{B-22}$$

$$\Gamma_Y(y) \triangleq \sum_{x\in\mathcal{X}}\sqrt{P_{X|Y}(x|y)}\Gamma(y,x). \tag{B-23}$$

此外，由 (B-20) 及 (B-21) 可推得

$$\sqrt{\hat{P}_X(x)\hat{P}_Y(y)} = \sqrt{P_X(x)P_Y(y)}\left[1+\frac{\tau}{2}\left(\frac{\Gamma_X(x)}{\sqrt{P_X(x)}}+\frac{\Gamma_Y(y)}{\sqrt{P_Y(y)}}\right)\right] + o(\tau) \tag{B-24}$$

以及

$$\frac{1}{\sqrt{\hat{P}_X(x)\hat{P}_Y(y)}} = \frac{1}{\sqrt{P_X(x)P_Y(y)}}\left[1-\frac{\tau}{2}\left(\frac{\Gamma_X(x)}{\sqrt{P_X(x)}}+\frac{\Gamma_Y(y)}{\sqrt{P_Y(y)}}\right)\right] + o(\tau). \tag{B-25}$$

因此，根据 (4-15) 及 (B-20)–(B-25)，对任意 $(x,y)\in\mathcal{X}\times\mathcal{Y}$ 我们有

$$\begin{aligned}
&\hat{B}(y,x) - \tilde{B}(y,x) \qquad\qquad (\text{B-26})\\
&= \tau\Bigg(\frac{\sqrt{P_{XY}(x,y)}}{\sqrt{P_X(x)P_Y(y)}}\Gamma(y,x) - \frac{P_{XY}(x,y)+P_X(x)P_Y(y)}{2\sqrt{P_X(x)P_Y(y)}}\\
&\quad\cdot\left[\frac{1}{P_X(x)}\sum_{y'\in\mathcal{Y}}\sqrt{P_{XY}(x,y')}\Gamma(y',x) + \frac{1}{P_Y(y)}\sum_{x'\in\mathcal{X}}\sqrt{P_{XY}(x',y)}\Gamma(y,x')\right]\Bigg) + o(\tau)\\
&= \tau\Xi(y,x) + o(\tau)
\end{aligned}$$

$$= \sqrt{\epsilon}\,\Xi(y,x) + o\left(\sqrt{\epsilon}\right), \tag{B-27}$$

亦即 (4-17)。

B.4　引理 4.4 的证明

对给定 $\epsilon > 0$ 及 $t > 0$，定义 $\mathcal{N}(\epsilon)$ 的子集 $S_2^{(t)}(\epsilon)$ 为

$$S_2^{(t)}(\epsilon) \triangleq \left\{ \hat{P}_{XY} :\ \hat{P}_{XY} \in \mathcal{N}(\epsilon), \hat{P}_{XY} \leftrightarrow \Gamma,\ \sum_{i=1}^{k}\sum_{j=k+1}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\Xi + \Xi^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} \geqslant t \right\}, \tag{B-28}$$

其中 $\boldsymbol{\phi}_i$ 为矩阵 $\tilde{\mathbf{B}}$ 第 i 个右奇异向量，Ξ 的定义由 (4-16) 给出。我们有如下关于 $S_2^{(t)}(\epsilon)$ 的引理。

引理 B.1： 对任意 $t \in (0,2)$，我们有

$$\lim_{\epsilon\to 0^+} \epsilon^{-1} D\left(S_2^{(t)}(\epsilon) \,\middle\|\, P_{XY}\right) = \frac{t}{2\alpha_k}. \tag{B-29}$$

基于引理 B.1，我们给出 引理 4.4 的证明如下。首先对所有满足 $\hat{P}_{XY} \in \mathcal{N}(\epsilon)$ 的经验分布 $\hat{P}_{XY}$，由引理 4.3 可知

$$\hat{\mathbf{B}}^{\mathsf{T}}\hat{\mathbf{B}} = \tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}} + \sqrt{\epsilon}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\Xi + \Xi^{\mathsf{T}}\tilde{\mathbf{B}}\right) + o\left(\sqrt{\epsilon}\right). \tag{B-30}$$

利用引理 4.1 中微扰分析的结果，我们可以将学习误差表示为

$$\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 = \epsilon \sum_{i=1}^{k}\sum_{j=k+1}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\Xi + \Xi^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} + o(\epsilon). \tag{B-31}$$

因此，对任意 $t \in (0,1)$，存在 $\epsilon_0 > 0$ 使得对所有 $\epsilon \in (0, \epsilon_0)$，我们有

$$S_2^{(1+t)}(\epsilon) \subseteq S_1(\epsilon) \cap \mathcal{N}(\epsilon) \subseteq S_2^{(1-t)}(\epsilon). \tag{B-32}$$

由此推得

$$D\left(S_2^{(1-t)}(\epsilon) \,\middle\|\, P_{XY}\right) \leqslant D\left(S_1(\epsilon) \cap \mathcal{N}(\epsilon) \,\middle\|\, P_{XY}\right) \leqslant D\left(S_2^{(1+t)}(\epsilon) \,\middle\|\, P_{XY}\right). \tag{B-33}$$

由 (B-33) 的第一个不等式可知

$$\liminf_{\epsilon\to 0^+} \epsilon^{-1} D\left(S_1(\epsilon) \cap \mathcal{N}(\epsilon) \,\middle\|\, P_{XY}\right) \geqslant \liminf_{\epsilon\to 0^+} \epsilon^{-1} D\left(S_2^{(1-t)}(\epsilon) \,\middle\|\, P_{XY}\right) = \frac{1-t}{2\alpha_k}, \tag{B-34a}$$

其中等号由引理 B.1 推出。类似地，根据 (B-33) 的第二个不等式可知

$$\limsup_{\epsilon\to 0^+}\epsilon^{-1}D\left(\mathcal{S}_1(\epsilon)\cap\mathcal{N}(\epsilon)\,\middle\|\,P_{XY}\right)\leqslant\limsup_{\epsilon\to 0^+}\epsilon^{-1}D\left(\mathcal{S}_2^{(1+t)}(\epsilon)\,\middle\|\,P_{XY}\right)=\frac{1+t}{2\alpha_k}.\tag{B-34b}$$

注意到由于 t 可以任意接近 0，必然有

$$\lim_{\epsilon\to 0^+}\epsilon^{-1}D\left(\mathcal{S}_1(\epsilon)\cap\mathcal{N}(\epsilon)\,\middle\|\,P_{XY}\right)=\frac{1}{2\alpha_k}\tag{B-35}$$

成立。

接下来只需证明引理 B.1。

证明 (引理 B.1 的证明) 由于集合 $\mathcal{S}_2^{(t)}(\epsilon)$ 为闭集，我们有

$$D\left(\mathcal{S}_2^{(t)}(\epsilon)\,\middle\|\,P_{XY}\right)=\inf_{\hat{P}_{XY}\in\mathcal{S}_2^{(t)}(\epsilon)}D(\hat{P}_{XY}\|P_{XY})=\min_{\hat{P}_{XY}\in\mathcal{S}_2^{(t)}(\epsilon)}D(\hat{P}_{XY}\|P_{XY}).$$

其次，对所有满足 $\hat{P}_{XY}\leftrightarrow\Gamma$ 及 $\hat{P}_{XY}\in\mathcal{S}_2^{(t)}(\epsilon)\subset\mathcal{N}(\epsilon)$ 的 Γ，由 (B-14) 可知对所有的 $(x,y)\in\mathcal{X}\times\mathcal{Y}$，$\Gamma(y,x)$ 均有界。因此，根据 K-L 散度的二阶展开 (参考命题 2.1) 有

$$\begin{aligned}D(\hat{P}_{XY}\|P_{XY})&=\frac{1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\frac{\left[\hat{P}_{XY}(x,y)-P_{XY}(x,y)\right]^2}{P_{XY}(x,y)}+o(\epsilon)\\&=\frac{\epsilon}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\Gamma^2(y,x)+o(\epsilon)=\frac{\epsilon}{2}\left\|\boldsymbol{\Gamma}\right\|_{\mathrm{F}}^2+o(\epsilon).\end{aligned}\tag{B-36}$$

另外，因为 $\hat{P}_{XY}$ 及 P_{XY} 均为概率分布，由 (4-15) 可知

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\sqrt{P_{XY}(x,y)}\Gamma(y,x)=0.\tag{B-37}$$

因此，对 (B-29) 的求解可转化为解如下优化问题：

$$\underset{\boldsymbol{\Gamma}}{\text{minimize}}\quad\left\|\boldsymbol{\Gamma}\right\|_{\mathrm{F}}^2\tag{B-38a}$$

$$\text{subject to}\quad\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\geqslant t,\tag{B-38b}$$

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\sqrt{P_{XY}(x,y)}\Gamma(y,x)=0.\tag{B-38c}$$

虽然上述优化问题中并未考虑限制条件 $\hat{P}_{XY} \in \mathcal{N}(\epsilon)$，但我们很快可从推导中看到，最优的 $\boldsymbol{\Gamma}$ 自动满足该条件。注意到因为该优化问题中目标函数及不等式约束条件 (B-38b) 均为二次的，其最优解可通过求解如下优化问题求得：

$$\begin{aligned} \underset{\boldsymbol{\Gamma}}{\text{maximize}} \quad & \sum_{i=1}^{k}\sum_{j=k+1}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2} \\ \text{subject to} \quad & \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \leqslant 1, \ \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \sqrt{P_{XY}(x,y)}\Gamma(y,x)=0, \end{aligned} \tag{B-39}$$

其中我们对调了目标函数及不等式约束中的二次函数。进一步地，易知优化问题 (B-39) 等价于其除掉等式约束所得优化问题，即

$$\underset{\boldsymbol{\Gamma}}{\text{maximize}} \quad \sum_{i=1}^{k}\sum_{j=k+1}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2} \tag{B-40a}$$

$$\text{subject to} \quad \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \leqslant 1. \tag{B-40b}$$

实际上，令 $\boldsymbol{\Gamma}^*$ 为 (B-40) 的最优解，定义 $c \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \sqrt{P_{XY}(x,y)}\Gamma^*(y,x)$ 及 $z(x,y) \triangleq \Gamma^*(y,x) - c\sqrt{P_{XY}(x,y)}$，则有

$$1 = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \left[\Gamma^*(y,x)\right]^2 = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} z^2(x,y) + c^2,$$

故 $|c| \leqslant 1$。

下面对 $|c|$ 取值进行讨论。若 $|c|=1$，可得 $\Gamma^*(y,x) = \pm\sqrt{P_{XY}(x,y)}$，于是由 (4-16) 可推出 $\Xi(y,x) = \mp\sqrt{P_X(x)P_Y(y)}$。因此得出 $\boldsymbol{\Xi} = \mp\boldsymbol{\psi}_0\boldsymbol{\phi}_0^{\mathsf{T}}$，其中 $\boldsymbol{\psi}_0$ 为第 y 个元素为 $\sqrt{P_Y(y)}$ 的 $|\mathcal{Y}|$ 维向量；类似地，$\boldsymbol{\phi}_0 \in \mathbb{R}^{|\mathcal{X}|}$ 的第 x 个元素定义为 $\sqrt{P_X(x)}$。于是由 $\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} = \mp\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\psi}_0\boldsymbol{\psi}_0^{\mathsf{T}} = \mathbf{O}$ 可知目标函数值为 0，与 $\boldsymbol{\Gamma}^*$ 最优的假设矛盾。

若 $0 < |c| < 1$, 构造矩阵 $\boldsymbol{\Gamma}'$ 使其元素为 $\Gamma'(y,x) = z(x,y)/\sqrt{1-c^2}$，则易得 $\left\|\boldsymbol{\Gamma}'\right\|_{\mathrm{F}}^2 = 1$ 且对优化问题 (B-39)，当优化变量 $\boldsymbol{\Gamma}$ 取 $\boldsymbol{\Gamma}'$ 时所对应的目标函数值为取 $\boldsymbol{\Gamma}^*$ 时所对应值的 $1/(1-c^2)$ 倍，与 $\boldsymbol{\Gamma}^*$ 最优的假设矛盾。

综上，必然有 $c=0$，因此优化问题 (B-40) 与 (B-39) 有相同的最优解。

此外，可验证对所有满足 $P_{XY}(x',y')=0$ 的 $(x',y') \in \mathcal{X}\times\mathcal{Y}$，必然有 $\Gamma^*(y',x')=0$ 成立。如若不然，令 $\Gamma^*(y',x')=0$ 且对 $\boldsymbol{\Gamma}^*$ 进行常数倍缩放使其满足 $\left\|\boldsymbol{\Gamma}^*\right\|_{\mathrm{F}}^2 = 1$。根据 (4-16)，(B-40) 的目标函数值将增加，这与 $\boldsymbol{\Gamma}^*$ 最优的假设矛盾。因此，最优解 $\boldsymbol{\Gamma}^*$ 满足定义式 (4-15)。

为简化目标函数(B-40a)，我们引入向量化操作 vec(·)，将(B-40a)表示为

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\\
&\quad=\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\left(\tilde{\mathbf{B}}\boldsymbol{\phi}_i\right)^{\mathsf{T}}\boldsymbol{\Xi}\boldsymbol{\phi}_j+\left(\tilde{\mathbf{B}}\boldsymbol{\phi}_j\right)^{\mathsf{T}}\boldsymbol{\Xi}\boldsymbol{\phi}_i\right]^2}{\sigma_i^2-\sigma_j^2} &\text{(B-41a)}\\
&\quad=\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left(\operatorname{tr}\left\{\left(\tilde{\mathbf{B}}\boldsymbol{\phi}_i\boldsymbol{\phi}_j^{\mathsf{T}}+\tilde{\mathbf{B}}\boldsymbol{\phi}_j\boldsymbol{\phi}_i^{\mathsf{T}}\right)^{\mathsf{T}}\boldsymbol{\Xi}\right\}\right)^2}{\sigma_i^2-\sigma_j^2} &\text{(B-41b)}\\
&\quad=\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\operatorname{vec}^{\mathsf{T}}\left(\tilde{\mathbf{B}}\boldsymbol{\phi}_i\boldsymbol{\phi}_j^{\mathsf{T}}+\tilde{\mathbf{B}}\boldsymbol{\phi}_j\boldsymbol{\phi}_i^{\mathsf{T}}\right)\operatorname{vec}(\boldsymbol{\Xi})\right]^2}{\sigma_i^2-\sigma_j^2} &\text{(B-41c)}\\
&\quad=\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\theta}_{ij}^{\mathsf{T}}\operatorname{vec}(\boldsymbol{\Xi})\right]^2}{\sigma_i^2-\sigma_j^2} &\text{(B-41d)}\\
&\quad=\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Xi})\left(\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}\right)\operatorname{vec}(\boldsymbol{\Xi}), &\text{(B-41e)}
\end{aligned}$$

其中 (B-41b)–(B-41c) 使用了矩阵迹的性质

$$\mathbf{u}^{\mathsf{T}}\mathbf{M}\mathbf{v}=\operatorname{tr}\left\{\mathbf{u}^{\mathsf{T}}\mathbf{M}\mathbf{v}\right\}=\operatorname{tr}\left\{\mathbf{v}\mathbf{u}^{\mathsf{T}}\mathbf{M}\right\}=\operatorname{vec}^{\mathsf{T}}\left(\mathbf{u}\mathbf{v}^{\mathsf{T}}\right)\operatorname{vec}(\mathbf{M}),$$

式 (B-41d) 基于恒等关系

$$\operatorname{vec}\left(\tilde{\mathbf{B}}\boldsymbol{\phi}_i\boldsymbol{\phi}_j^{\mathsf{T}}+\tilde{\mathbf{B}}\boldsymbol{\phi}_j\boldsymbol{\phi}_i^{\mathsf{T}}\right)=\boldsymbol{\phi}_j\otimes(\tilde{\mathbf{B}}\boldsymbol{\phi}_i)+\boldsymbol{\phi}_i\otimes(\tilde{\mathbf{B}}\boldsymbol{\phi}_j)=\boldsymbol{\theta}_{ij}.$$

在此基础上，由 (4-16) 及 (4-11) 可得 $\operatorname{vec}(\boldsymbol{\Xi})=\mathbf{L}\operatorname{vec}(\boldsymbol{\Gamma})$，于是 (B-41e) 可化简为

$$\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Gamma})\mathbf{L}^{\mathsf{T}}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}\right)\mathbf{L}\operatorname{vec}(\boldsymbol{\Gamma})=\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Gamma})\mathbf{G}_k\operatorname{vec}(\boldsymbol{\Gamma}). \tag{B-42}$$

根据 $\|\operatorname{vec}(\boldsymbol{\Gamma})\|=\|\boldsymbol{\Gamma}\|_{\mathrm{F}}$，约束 (B-40) 等价于 $\|\operatorname{vec}(\boldsymbol{\Gamma})\|\leqslant 1$，因此 (B-42) 的最大值为 $\mathbf{G}_k$ 的最大奇异值 α_k，即优化问题 (B-40) 及 (B-39) 的目标函数最优值。由此可知，原优化问题 (B-38) 的最优解为 $\sqrt{\frac{t}{\alpha_k}}\boldsymbol{\Gamma}^*$，相应的目标函数最优值为 $t\alpha_k^{-1}$。令

$\hat{P}^*_{XY} \leftrightarrow \sqrt{\frac{t}{\alpha_k}}\Gamma^*$ 为对应的最优经验分布，则对充分小的 ϵ，有

$$D(\hat{P}^*_{XY}\|P_{XY}) = \frac{\epsilon t}{2\alpha_k} + o(\epsilon) < \frac{\epsilon}{\alpha_k},$$

其中不等号成立由于 $t \in (0,2)$.

由此得出 $\hat{P}^*_{XY} \in \mathcal{N}(\epsilon)$，故

$$\min_{\hat{P}_{XY}\in S_2^{(t)}(\epsilon)} D(\hat{P}_{XY}\|P_{XY}) = D(\hat{P}^*_{XY}\|P_{XY}) + o(\epsilon) = \frac{\epsilon t}{2\alpha_k} + o(\epsilon), \tag{B-43}$$

□

从而 (B-29) 成立。

B.5　定理 4.2 的证明

首先由 (4-22) 知存在 $\epsilon_1 > 0$ 使得对任意 $\epsilon \in (0,\epsilon_1)$ 有

$$D\left(S_1(\epsilon)\,\middle\|\, P_{XY}\right) = \frac{\epsilon}{2\alpha_k} + o(\epsilon) > \frac{\epsilon}{3\alpha_k}.$$

于是根据 Sanov 定理有

$$\begin{aligned}\mathbb{P}_n\left\{\left\|\tilde{\mathbf{B}}\mathbf{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\mathbf{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} &\leqslant (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\cdot D\left(S_1(\epsilon)\,\middle\|\, P_{XY}\right)\right)\\ &< (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-\frac{n\epsilon}{3\alpha_k}\right).\end{aligned} \tag{B-44}$$

因此只需选择 n 使其满足

$$(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-\frac{n\epsilon}{3\alpha_k}\right) < \delta,$$

即

$$n > \frac{1}{\mu}W_{-1}\left(\mu e^{\mu}\delta^{\frac{1}{|\mathcal{X}||\mathcal{Y}|}}\right) - 1, \tag{B-45}$$

其中 $\mu \triangleq -\frac{\epsilon}{3|\mathcal{X}||\mathcal{Y}|\alpha_k}$，且 $W_{-1}(\cdot)$ 表示下半支的 Lambert W 函数[95]。

注意到由于当 $x \to 0^-$ 时有 $W_{-1}(x) = \log(-x) + o(\log(-x))$，存在 $x_0 < 0$ 使得

$$W_{-1}(x) > \frac{4}{3}\log(-x), \quad \forall\, x \in (x_0, 0). \tag{B-46}$$

因此，令 $\text{let}\ \epsilon_2 \triangleq \min\left\{3|\mathcal{X}||\mathcal{Y}|\alpha_k|x_0|, \frac{\alpha_k}{3|\mathcal{X}||\mathcal{Y}|e}\right\}$，则对任意 $\epsilon \in (0,\epsilon_2)$ 有

$$\left|\mu e^{\mu}\delta^{\frac{1}{|\mathcal{X}||\mathcal{Y}|}}\right| < |\mu| = \frac{\epsilon}{3|\mathcal{X}||\mathcal{Y}|\alpha_k} < |x_0|,$$

以及

$$\frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log(3|\mathcal{X}||\mathcal{Y}|) = \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{\alpha_k}{\epsilon} + \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{3|\mathcal{X}||\mathcal{Y}|\epsilon}{\alpha_k} \tag{B-47a}$$

$$< \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{\alpha_k}{\epsilon} - \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon} \tag{B-47b}$$

$$< \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{\alpha_k}{\epsilon} - 12|\mathcal{X}|^2|\mathcal{Y}|^2 e, \tag{B-47c}$$

其中 (B-47) 中的不等式基于 $\epsilon < \epsilon_2 < \frac{\alpha_k}{3|\mathcal{X}||\mathcal{Y}|e}$。

故得

$$\begin{aligned}
&\frac{1}{\mu}W_{-1}\left(\mu e^{\mu}\delta^{\frac{1}{|\mathcal{X}||\mathcal{Y}|}}\right) - 1 \\
&\quad < \frac{4}{3\mu}\log\left(-\mu e^{\mu}\delta^{\frac{1}{|\mathcal{X}||\mathcal{Y}|}}\right) - 1 && \text{(B-48a)}\\
&\quad = \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{\alpha_k}{\epsilon} + \frac{4\alpha_k}{\epsilon}\log\frac{1}{\delta} + \frac{4|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log(3|\mathcal{X}||\mathcal{Y}|) + \frac{1}{3} && \text{(B-48b)}\\
&\quad < \frac{8|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{\alpha_k}{\epsilon} + \frac{4\alpha_k}{\epsilon}\log\frac{1}{\delta} - 12|\mathcal{X}|^2|\mathcal{Y}|^2 e + \frac{1}{3} && \text{(B-48c)}\\
&\quad < \frac{8|\mathcal{X}||\mathcal{Y}|\alpha_k}{\epsilon}\log\frac{\alpha_k}{\epsilon} + \frac{4\alpha_k}{\epsilon}\log\frac{1}{\delta} && \text{(B-48d)}\\
&\quad = N(\epsilon,\delta), && \text{(B-48e)}
\end{aligned}$$

其中 (B-48a) 基于 (B-46), (B-48c) 由 (B-47) 推出。

最后，令 $\epsilon_0 \triangleq \min\{\epsilon_1, \epsilon_2\}$，则对任意 $\epsilon \in (0, \epsilon_0)$ 及 $n \geqslant N(\epsilon,\delta)$，由 (B-44) 与 (B-48) 得

$$\mathbb{P}_n\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} < (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-\frac{n\epsilon}{3\alpha_k}\right) < \delta,$$

从而定理得证。

B.6　推论 4.1 的证明

首先注意到 (4-10) 可化简为

$$\mathbf{G}_k = \sum_{i=1}^{d-1}\frac{1}{\sigma_i^2}\mathbf{L}^{\mathsf{T}}\boldsymbol{\theta}_{id}\boldsymbol{\theta}_{id}^{\mathsf{T}}\mathbf{L} = \sum_{i=1}^{d-1}\frac{1}{\sigma_i^2}\left(\mathbf{L}^{\mathsf{T}}\boldsymbol{\theta}_{id}\right)\left(\mathbf{L}^{\mathsf{T}}\boldsymbol{\theta}_{id}\right)^{\mathsf{T}}, \tag{B-49}$$

其中

$$\boldsymbol{\theta}_{id} = \boldsymbol{\phi}_d \otimes \left(\tilde{\mathbf{B}}\boldsymbol{\phi}_i\right) = \sigma_i(\boldsymbol{\phi}_d \otimes \boldsymbol{\psi}_i), \quad i = 1, \cdots, d-1,$$

且 $\boldsymbol{\psi}_i \in \mathbb{R}^{|\mathcal{Y}|}$ 表示 $\tilde{\mathbf{B}}$ 的第 i 个左奇异向量。由 $\sigma_{d-1} > 0 = \sigma_d$ 以及 $d = |\mathcal{X}| \leqslant |\mathcal{Y}|$ 可知 $\boldsymbol{\phi}_d$ 为奇异值 0 所对应的唯一右奇异向量，因此

$$\boldsymbol{\phi}_d = \left[\sqrt{P_X(1)}, \cdots, \sqrt{P_X(d)}\right]^{\mathsf{T}}. \tag{B-50}$$

从而由 (4-11) 可得 $\left(\mathbf{L}^{\mathsf{T}}\boldsymbol{\theta}_{id}\right)$ 的第 $[(x-1)|\mathcal{Y}| + y]$ 个元素为

$$\begin{aligned}
&\sum_{x'\in\mathcal{X},y'\in\mathcal{Y}} \sqrt{\frac{P_{XY}(x,y)}{P_X(x')P_Y(y')}}\left[\sigma_i\phi_d(x')\psi_i(y')\right] \\
&\quad\cdot\left(\delta_{xx'}\delta_{yy'} - \frac{1}{2}\left[\frac{\delta_{xx'}}{P_X(x')} + \frac{\delta_{yy'}}{P_Y(y')}\right]\cdot\left[P_{XY}(x',y') + P_X(x')P_Y(y')\right]\right) \\
&= \sigma_i\phi_d(x)\psi_i(y)\sqrt{\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}} - \frac{\sigma_i}{2}\sqrt{P_{XY}(x,y)} \\
&\quad\cdot\sum_{x'\in\mathcal{X},y'\in\mathcal{Y}}\left(\left[\frac{\delta_{xx'}}{P_X(x')} + \frac{\delta_{yy'}}{P_Y(y')}\right]\cdot\left[\tilde{B}(y',x') + 2\phi_d(x')\sqrt{P_Y(y')}\right]\psi_i(y')\phi_d(x')\right)
\end{aligned} \tag{B-51a}$$

$$\begin{aligned}
&= \sigma_i\phi_d(x)\psi_i(y)\sqrt{\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}} - \frac{\sigma_i}{2}\sqrt{P_{XY}(x,y)} \\
&\quad\cdot\left(\frac{\phi_d(x)}{P_X(x)}\sum_{y'\in\mathcal{Y}}\left[\tilde{B}(y',x) + 2\phi_d(x)\sqrt{P_Y(y')}\right]\psi_i(y')\right. \\
&\qquad\left. + \frac{\psi_i(y)}{P_Y(y)}\sum_{x'\in\mathcal{X}}\left[\tilde{B}(y,x') + 2\phi_d(x')\sqrt{P_Y(y)}\right]\phi_d(x')\right)
\end{aligned} \tag{B-51b}$$

$$= \sigma_i\phi_d(x)\psi_i(y)\sqrt{\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}} - \frac{\sigma_i}{2}\sqrt{P_{XY}(x,y)}\cdot\left[\frac{\sigma_i\phi_i(x)\phi_d(x)}{P_X(x)} + \frac{2\psi_i(y)}{\sqrt{P_Y(y)}}\right] \tag{B-51c}$$

$$= -\frac{\sigma_i^2}{2}\sqrt{P_{Y|X}(y|x)}\phi_i(x), \tag{B-51d}$$

其中 $\phi_i(x)$ 与 $\psi_j(y)$ 分别表示 $\boldsymbol{\phi}_i$ 的第 x 个元素及 $\boldsymbol{\psi}_j$ 的第 y 个元素，式 (B-51a) 的推导基于

$$\frac{P_{XY}(x',y') + P_X(x')P_Y(y')}{\sqrt{P_X(x')P_Y(y')}} = \tilde{B}(y',x') + 2\phi_d(x')\sqrt{P_Y(y')}.$$

此外，式 (B-51c) 的推导基于

$$\sum_{x'\in\mathcal{X}}\tilde{B}(y,x')\phi_d(x') = \sigma_d\psi_d(y) = 0, \tag{B-52a}$$

$$\sum_{x'\in\mathcal{X}}[\phi_d(x')]^2=\|\boldsymbol{\phi}_d\|^2=1, \tag{B-52b}$$

以及对给定 $1\leqslant i\leqslant d-1$，有

$$\sum_{y'\in\mathcal{Y}}\tilde{B}(y',x)\psi_i(y')=\sigma_i\phi_i(x), \tag{B-53a}$$

$$\sum_{y'\in\mathcal{Y}}\sqrt{P_Y(y')}\psi_i(y')=0, \tag{B-53b}$$

其中 (B-53b) 成立由于向量 $\left[\sqrt{P_Y(1)},\cdots,\sqrt{P_Y(|\mathcal{Y}|)}\right]^{\mathsf{T}}\in\mathbb{R}^{|\mathcal{Y}|}$ 为矩阵 $\tilde{\mathbf{B}}$ 对应于奇异值 0 的左奇异向量。

故由 (B-51) 可得

$$\mathbf{L}^{\mathsf{T}}\boldsymbol{\theta}_{id}=-\frac{\sigma_i^2}{2}\mathbf{M}\boldsymbol{\phi}_i,$$

其中 $\mathbf{M}\in\mathbb{R}^{(|\mathcal{X}|\cdot|\mathcal{Y}|)\times|\mathcal{X}|}$ 为块对角矩阵，其定义为

$$\mathbf{M}\triangleq\begin{bmatrix}\boldsymbol{\nu}_1 & \mathbf{0}_{|\mathcal{Y}|} & \cdots & \mathbf{0}_{|\mathcal{Y}|}\\ \mathbf{0}_{|\mathcal{Y}|} & \boldsymbol{\nu}_2 & \cdots & \mathbf{0}_{|\mathcal{Y}|}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0}_{|\mathcal{Y}|} & \mathbf{0}_{|\mathcal{Y}|} & \cdots & \boldsymbol{\nu}_{|\mathcal{X}|}\end{bmatrix}, \tag{B-54a}$$

式中 $\mathbf{0}_{|\mathcal{Y}|}$ 为 $\mathbb{R}^{|\mathcal{Y}|}$ 中的零向量，且对所有 $x\in\mathcal{X}$，$\boldsymbol{\nu}_x$ 定义为 $|\mathcal{Y}|$ 维向量：

$$\boldsymbol{\nu}_x=\left[\sqrt{P_{Y|X}(1|x)},\cdots,\sqrt{P_{Y|X}\left(|\mathcal{Y}||x\right)}\right]^{\mathsf{T}}. \tag{B-54b}$$

因此，由 (B-49) 推出

$$\mathbf{G}_k=\frac{1}{4}\sum_{i=1}^{d-1}\sigma_i^2(\mathbf{M}\boldsymbol{\phi}_i)(\mathbf{M}\boldsymbol{\phi}_i)^{\mathsf{T}}, \tag{B-55}$$

从中可得 $\mathbf{G}_k$ 的特征值分解。实际上，由 $\mathbf{M}^{\mathsf{T}}\mathbf{M}=\mathbf{I}_d$ 可知 $\langle\mathbf{M}\boldsymbol{\phi}_i,\mathbf{M}\boldsymbol{\phi}_j\rangle=\langle\boldsymbol{\phi}_i,\boldsymbol{\phi}_j\rangle=\delta_{ij}$。因此，从 (B-55) 得出 $\mathbf{G}_k$ 的非零特征值为 $\sigma_i^2/4\,(i=1,\cdots,d-1)$，对应特征向量 $\mathbf{M}\boldsymbol{\phi}_i\,(i=1,\cdots,d-1)$。故 $\mathbf{G}_k$ 的最大特征值 (即最大奇异值) 为

$$\alpha_k=\|\mathbf{G}_k\|_{\mathrm{s}}=\frac{\sigma_1^2}{4},$$

其中 $\|\cdot\|_{\mathrm{s}}$ 表示矩阵的谱范数。

B.7　定理 4.3 的证明

该定理证明过程与定理 4.1 类似，只需将引理 4.1 的扰动分析结果改为 引理 4.2 的相应结果。首先注意到，$\mathcal{N}(\epsilon)$ 的定义可自然地拓展到 $\sigma_k = \sigma_{k+1}$ 的情形[①]：

$$\mathcal{N}(\epsilon) \triangleq \left\{ \hat{P}_{XY} :\ D(\hat{P}_{XY} \| P_{XY}) \leqslant \frac{\epsilon}{\beta_k} \right\}. \tag{B-56}$$

接着定义 $\mathcal{S}_3^{(t)}(\epsilon)$ 为关于 $\hat{P}_{XY}$ 的集合，使得由 (4-15) 给出的对应 Γ 满足

$$\sum_{i=1}^{l-1}\sum_{j=l}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} + \sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}} \frac{\left[\boldsymbol{\varphi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} \geqslant t, \tag{B-57}$$

其中 $\boldsymbol{\varphi}_i$ 的定义参见 (4-25)。类似于 引理 B.1，我们有如下结果。

引理 B.2： 当 $\sigma_k = \sigma_{k+1}$ 时，对任意 $t \in (0, 2)$，有

$$\lim_{\epsilon\to 0^+} \epsilon^{-1} D\left(\mathcal{S}_3^{(t)}(\epsilon) \,\middle\|\, P_{XY}\right) = \frac{t}{2\beta_k}. \tag{B-58}$$

证明 该引理证明过程与 引理 B.1 类似。根据 K-L 散度的二阶 Taylor 展开 (B-36)，极限 (B-58) 可通过解如下优化问题求得：

$$\begin{aligned}
\underset{\boldsymbol{\Gamma}}{\text{minimize}} \quad & \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \\
\text{subject to} \quad & \sum_{i=1}^{l-1}\sum_{j=l}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} + \sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}} \frac{\left[\boldsymbol{\varphi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} \geqslant t, \qquad \text{(B-59a)} \\
& \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \sqrt{P_{XY}(x,y)}\,\Gamma(y,x) = 0. \qquad \text{(B-59b)}
\end{aligned}$$

与 引理 B.1 的证明过程同理，(B-59) 的最优解可转化为求解

$$\begin{aligned}
\underset{\boldsymbol{\Gamma}}{\text{maximize}} \quad & \sum_{i=1}^{l-1}\sum_{j=l}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} + \sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}} \frac{\left[\boldsymbol{\varphi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi} + \boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} \qquad \text{(B-60a)} \\
\text{subject to} \quad & \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 \leqslant 1, \qquad \text{(B-60b)}
\end{aligned}$$

其中我们对调了目标函数及不等式约束中的二次函数，并去掉了等式约束。

① 易验证当 $\sigma_k > \sigma_{k+1}$ 时有 $\beta_k = \alpha_k$，因此 (B-56) 可视为 (4-14) 的推广。

类似于 (B-41)，目标函数 (B-60a) 可表示为

$$\begin{aligned}
&\sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}+\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left[\boldsymbol{\varphi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\\
&\quad=\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Xi})\left(\sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}+\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\boldsymbol{\vartheta}_{ij}\boldsymbol{\vartheta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}\right)\operatorname{vec}(\boldsymbol{\Xi})\\
&\quad=\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Gamma})\mathbf{L}^{\mathsf{T}}\left(\sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}+\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\boldsymbol{\vartheta}_{ij}\boldsymbol{\vartheta}_{ij}^{\mathsf{T}}}{\sigma_i^2-\sigma_j^2}\right)\mathbf{L}\operatorname{vec}(\boldsymbol{\Gamma})\\
&\quad=\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Gamma})\mathbf{J}_k(\boldsymbol{\Gamma})\operatorname{vec}(\boldsymbol{\Gamma}),
\end{aligned}$$

其中第二个等号成立因为 $\operatorname{vec}(\boldsymbol{\Xi})=\mathbf{L}\operatorname{vec}(\boldsymbol{\Gamma})$。因此优化问题 (B-60) 可等价表示为 (4-26)，从而其最优解为 β_k。注意到对 $\sigma_k>\sigma_{k+1}$ 的特例，可不妨取 $\boldsymbol{\varphi}_i=\boldsymbol{\phi}_i$ ($i=l,\cdots,k$)。易知此时 (B-60) 的最优值为 α_k，即当 $\sigma_k>\sigma_{k+1}$ 时有 $\beta_k=\alpha_k$。

最后，与 引理 B.1 证明过程同理，可得 (B-59) 最优值为 t/β_k ，从而

$$D\left(\mathcal{S}_3^{(t)}(\epsilon)\,\middle\|\,P_{XY}\right)=\frac{\epsilon t}{2\beta_k}+o(\epsilon),$$

由此可推出 (B-58)。 □

此外，由引理 4.2 及引理 4.3 可得经验分布 $\hat{P}_{XY}\in\mathcal{N}(\epsilon)$ 所造成的学习误差为

$$\begin{aligned}
\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\hat{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2=\epsilon\sum_{i=1}^{l-1}\sum_{j=l}^{d}&\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\\
&+\epsilon\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left[\boldsymbol{\varphi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\Xi}+\boldsymbol{\Xi}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}+o(\epsilon).
\end{aligned}\tag{B-61}$$

因此任取 $t\in(0,1)$，存在 $\epsilon_0>0$ 使得对任意 $\epsilon\in(0,\epsilon_0)$，有

$$\mathcal{S}_3^{(1+t)}(\epsilon)\subseteq\mathcal{S}_1(\epsilon)\cap\mathcal{N}(\epsilon)\subseteq\mathcal{S}_3^{(1-t)}(\epsilon).$$

类比于 (B-33)–(B-35) 的推导，可得

$$\lim_{\epsilon\to0^+}\epsilon^{-1}D\left(\mathcal{S}_1^{(t)}(\epsilon)\,\middle\|\,P_{XY}\right)=\frac{2}{\beta_k}.$$

最后，与 定理 4.1 同理，可得 (4-27)。

B.8　推论 4.2 的证明

首先，我们介绍两个相关的引理。

引理 B.3： 设 P_{XY} 由推论 4.2 所定义，则其对应典型相关矩阵 $\tilde{\mathbf{B}}$ 的奇异值为

$$\sigma_1 = \cdots = \sigma_{d-1} = \frac{|p_1 - p_2|\sqrt{d}}{\sqrt{p_1 + (d-1)p_2}}, \quad \sigma_d = 0. \tag{B-62}$$

此外，对所有满足

$$\langle \boldsymbol{\phi}, \mathbf{1}_d \rangle = 0 \quad \text{以及} \quad \|\boldsymbol{\phi}\| = 1 \tag{B-63}$$

的 $\boldsymbol{\phi} = [\phi(1), \cdots, \phi(d)]^\mathsf{T} \in \mathbb{R}^d$，其对应的 $\boldsymbol{\psi} \triangleq \sigma_1^{-1}\tilde{\mathbf{B}}^\mathsf{T}\boldsymbol{\phi} = [\psi(1), \cdots, \psi(|\mathcal{Y}|)]^\mathsf{T} \in \mathbb{R}^{|\mathcal{Y}|}$ 满足

$$\psi(y) = \begin{cases} \operatorname{sgn}(p_1 - p_2)\phi(y) & \text{若 } y \leqslant d \\ 0 & \text{其它情况,} \end{cases} \tag{B-64}$$

其中 $\mathbf{1}_d$ 表示 $\mathbb{R}^d$ 中所有元素均为 1 的向量。

证明 由 P_{XY} 定义可得

$$P_X(x) = \frac{1}{d}, \quad \forall\, x \in \mathcal{X},$$

以及

$$P_Y(y) = \begin{cases} p_1 + (d-1)p_2 & \text{若 } y \leqslant d \\ d\,p_2 & \text{其它情况.} \end{cases}$$

从而由定义 (6-21) 可得

$$\tilde{\mathbf{B}} = \frac{(p_1 - p_2)\sqrt{d}}{\sqrt{p_1 + (d-1)p_2}} \begin{bmatrix} \mathbf{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^\mathsf{T} \\ \mathbf{O}_{|\mathcal{Y}|-d,d} \end{bmatrix}, \tag{B-65}$$

其中 $\mathbf{O}_{|\mathcal{Y}|-d,d}$ 表示 $\mathbb{R}^{(|\mathcal{Y}|-d)\times d}$ 中的零矩阵。故可知

$$\tilde{\mathbf{B}}^\mathsf{T}\tilde{\mathbf{B}} = \frac{(p_1 - p_2)^2 d}{p_1 + (d-1)p_2}\left(\mathbf{I}_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\mathsf{T}\right).$$

由于矩阵

$$\mathbf{I}_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\mathsf{T}$$

特征值为 $\lambda_1 = \cdots = \lambda_{d-1} = 1$ 及 $\lambda_d = 0$，可得矩阵 $\tilde{\mathbf{B}}$ 的奇异值 $\sigma_1, \cdots, \sigma_d$ 如 (B-62) 所示，因此 (B-65) 可表示为

$$\tilde{\mathbf{B}} = \operatorname{sgn}(p_1 - p_2) \cdot \sigma_1 \begin{bmatrix} \mathbf{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^{\mathsf{T}} \\ \mathbf{O}_{|\mathcal{Y}|-d,d} \end{bmatrix}. \tag{B-66}$$

因此，对所有满足 (B-63) $\boldsymbol{\phi} \in \mathbb{R}^d$，有

$$\boldsymbol{\psi} = \sigma_1^{-1}\tilde{\mathbf{B}}\boldsymbol{\phi} = \operatorname{sgn}(p_1 - p_2) \begin{bmatrix} \mathbf{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^{\mathsf{T}} \\ \mathbf{O}_{|\mathcal{Y}|-d,d} \end{bmatrix} \boldsymbol{\phi} = \operatorname{sgn}(p_1 - p_2) \begin{bmatrix} \boldsymbol{\phi} \\ \mathbf{0}_{|\mathcal{Y}|-d} \end{bmatrix},$$

其中 $\mathbf{0}_{|\mathcal{Y}|-d}$ 为 $\mathbb{R}^{|\mathcal{Y}|-d}$ 中的零向量。 □

引理 B.4： 对所有满足 (B-63) 的 $\boldsymbol{\phi} = [\phi(1), \cdots, \phi(d)]^{\mathsf{T}} \in \mathbb{R}^d$，我们有

$$\phi^2(1) \leqslant \frac{d-1}{d},$$

式中等号成立当且仅当 $\boldsymbol{\phi} = \pm\boldsymbol{\phi}'$，其中

$$\boldsymbol{\phi}' \triangleq \frac{1}{\sqrt{(d-1)d}} [d-1, -1, \cdots, -1]^{\mathsf{T}} \in \mathbb{R}^d. \tag{B-67}$$

证明 由 $\langle \boldsymbol{\phi}, \mathbf{1}_d \rangle = 0$ 可得

$$\sum_{i=1}^{d} \phi(i) = 0.$$

由此得出

$$\phi^2(1) = \left[\sum_{i=2}^{d} \phi(i)\right]^2 \leqslant (d-1) \sum_{i=2}^{d} \phi^2(i) = (d-1)\left[\|\boldsymbol{\phi}\|^2 - \phi^2(1)\right],$$

其中不等式利用了算术平均不超过平方平均的事实。故

$$\phi^2(1) \leqslant \frac{d-1}{d}\|\boldsymbol{\phi}\|^2 = \frac{d-1}{d},$$

其中不等号取得等号当且仅当

$$\phi(2) = \phi(3) = \cdots = \phi(d). \tag{B-68}$$

因此，由 (B-63) 及 (B-68) 可知 $\boldsymbol{\phi} = \pm\boldsymbol{\phi}'$。 □

基于前述结论，推论 4.2 可证明如下。

证明 (推论 4.2 的证明) 由引理 B.3 可得 $\sigma_1 = \cdots = \sigma_{d-1} > \sigma_d = 0$，从而对任意 $1 \leqslant k \leqslant d-1$ 有 $\mathcal{I}_k = [d-1]$，由此推出 $l = \min \mathcal{I}_k = 1$ 以及 $\mathcal{I}_k^{\mathsf{c}} = \{d\}$。因此，根据 (4-24) 有

$$\mathbf{J}_k(\boldsymbol{\Gamma}) = \mathbf{L}^{\mathsf{T}} \left(\sum_{i=1}^{k} \frac{\boldsymbol{\vartheta}_{id} \boldsymbol{\vartheta}_{id}^{\mathsf{T}}}{\sigma_1^2} \right) \mathbf{L} = \frac{1}{\sigma_1^2} \sum_{i=1}^{k} \left(\mathbf{L}^{\mathsf{T}} \boldsymbol{\vartheta}_{id} \right) \left(\mathbf{L}^{\mathsf{T}} \boldsymbol{\vartheta}_{id} \right)^{\mathsf{T}}. \tag{B-69}$$

此外，与 (B-51) 同理，可得

$$\mathbf{L}^{\mathsf{T}} \boldsymbol{\vartheta}_{id} = -\frac{\sigma_1^2}{2} \mathbf{M} \boldsymbol{\varphi}_i, \quad 1 \leqslant i \leqslant k, \tag{B-70}$$

从而

$$\mathbf{J}_k(\boldsymbol{\Gamma}) = \frac{\sigma_1^2}{4} \sum_{i=1}^{k} (\mathbf{M} \boldsymbol{\varphi}_i)(\mathbf{M} \boldsymbol{\varphi}_i)^{\mathsf{T}}. \tag{B-71}$$

注意到由于 $\left\langle \mathbf{M} \boldsymbol{\varphi}_i, \mathbf{M} \boldsymbol{\varphi}_j \right\rangle = \delta_{ij}$，(B-71) 给出了 $\mathbf{G}_k$ 的特征值分解。故由定理 4.3 可推出

$$\beta_k \leqslant \left\| \mathbf{J}_k(\boldsymbol{\Gamma}) \right\|_{\mathrm{s}} = \frac{\sigma_1^4}{4}.$$

为证明上式中不等号可取得等号，只需构造 $\boldsymbol{\Gamma}$ 使其满足 $\|\boldsymbol{\Gamma}\|_{\mathsf{F}} \leqslant 1$ 以及

$$\operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Gamma})\, \mathbf{J}_k(\boldsymbol{\Gamma}) \operatorname{vec}(\boldsymbol{\Gamma}) = \left\| \mathbf{J}_k(\boldsymbol{\Gamma}) \right\|_{\mathrm{s}}. \tag{B-72}$$

实际上，对于 (B-67) 中的 $\boldsymbol{\phi}'$，构造 $\boldsymbol{\Gamma}$ 为

$$\Gamma(y, x) = \sqrt{P_{Y|X}(y|x)} \phi'(x) \tag{B-73}$$

则有 $\boldsymbol{\varphi}_1 = \pm \boldsymbol{\phi}'$ 以及 $\operatorname{vec}(\boldsymbol{\Gamma}) = \mathbf{M} \boldsymbol{\phi}' = \pm \mathbf{M} \boldsymbol{\varphi}_1$，从而 (B-72) 成立。

为说明该结论，首先注意到由 (B-73) 可得 $\|\boldsymbol{\Gamma}\|_{\mathsf{F}} = 1$，

$$\sum_{y' \in \mathcal{Y}} \sqrt{P_{XY}(x, y')} \Gamma(y', x) = \sqrt{P_X(x)} \phi'(x)$$

以及

$$\sum_{x' \in \mathcal{X}} \sqrt{P_{XY}(x', y)} \Gamma(y, x') = \sqrt{P_Y(y)} \sum_{x' \in \mathcal{X}} \frac{P_{XY}(x', y)}{\sqrt{P_X(x') P_Y(y)}} \phi'(x') = \sigma_1 \sqrt{P_Y(y)} \psi'(y),$$

其中 $\boldsymbol{\psi}' = \left[\psi'(1), \cdots, \psi'(|\mathcal{Y}|) \right]^{\mathsf{T}} \triangleq \sigma_1^{-1} \tilde{\mathbf{B}} \boldsymbol{\phi}'$.

因此，由 (4-16) 可得

$$\begin{aligned}\Xi(y,x) &= \frac{\sqrt{P_{XY}(x,y)}}{\sqrt{P_X(x)P_Y(y)}}\Gamma(y,x) - \frac{P_{XY}(x,y)+P_X(x)P_Y(y)}{2\sqrt{P_X(x)P_Y(y)}} \\ &\quad \cdot \left[\frac{1}{P_X(x)}\sum_{y'\in\mathcal{Y}}\sqrt{P_{XY}(x,y')}\Gamma(y',x) + \frac{1}{P_Y(y)}\sum_{x'\in\mathcal{X}}\sqrt{P_{XY}(x',y)}\Gamma(y,x')\right] \\ &= \frac{P_{XY}(x,y)}{\sqrt{P_X(x)P_Y(y)}}\cdot\frac{\phi'(x)}{\sqrt{P_X(x)}} - \frac{P_{XY}(x,y)+P_X(x)P_Y(y)}{2\sqrt{P_X(x)P_Y(y)}}\left[\frac{\phi'(x)}{\sqrt{P_X(x)}} + \frac{\sigma_1\psi'(y)}{\sqrt{P_Y(y)}}\right] \\ &= \frac{1}{2}\tilde{B}(y,x)\cdot\left[\frac{\phi'(x)}{\sqrt{P_X(x)}} - \frac{\sigma_1\psi'(y)}{\sqrt{P_Y(y)}}\right] - \sigma_1\sqrt{P_X(x)}\psi'(y). \end{aligned} \tag{B-74}$$

此外由于 $\mathcal{I}_k = [d-1]$，根据 (4-25) 可知 $\boldsymbol{\varphi}_1$ 为如下优化问题的解：

$$\begin{aligned} \underset{\boldsymbol{\phi}}{\text{maximize}} \quad & \boldsymbol{\phi}^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\Xi\right)\boldsymbol{\phi} \\ \text{subject to} \quad & \langle\boldsymbol{\phi},\boldsymbol{\phi}_d\rangle = 0, \quad \|\boldsymbol{\phi}\| = 1, \end{aligned} \tag{B-75}$$

其中 $\boldsymbol{\phi}_d$ 为 $\tilde{\mathbf{B}}$ 的第 d 个右奇异向量。由 $\sigma_{d-1} > 0 = \sigma_d$ 及 $d = |\mathcal{X}| \leqslant |\mathcal{Y}|$，可知

$$\boldsymbol{\phi}_d = \left[\sqrt{P_X(1)},\cdots,\sqrt{P_X(d)}\right]^{\mathsf{T}} = \frac{1}{\sqrt{d}}\mathbf{1}_d,$$

因此 $\langle\boldsymbol{\phi},\boldsymbol{\phi}_d\rangle = 0$ 等价于 $\langle\boldsymbol{\phi},\mathbf{1}_d\rangle = 0$.

于是对任意满足 (B-75) 约束的 $\boldsymbol{\phi}$，(B-75) 的目标函数可表为

$$\boldsymbol{\phi}^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\Xi\right)\boldsymbol{\phi} = \sigma_1\boldsymbol{\psi}^{\mathsf{T}}\Xi\boldsymbol{\phi} \tag{B-76a}$$

$$= \frac{\sigma_1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\tilde{B}(y,x)\cdot\left[\frac{\phi'(x)}{\sqrt{P_X(x)}} - \frac{\sigma_1\psi'(y)}{\sqrt{P_Y(y)}}\right]\phi(x)\psi(y)$$

$$- \sigma_1\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\sqrt{P_X(x)}\psi'(y)\phi(x)\psi(y) \tag{B-76b}$$

$$= \frac{\sigma_1}{2}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\tilde{B}(y,x)\cdot\left[\frac{\phi'(x)}{\sqrt{P_X(x)}} - \frac{\sigma_1\psi'(y)}{\sqrt{P_Y(y)}}\right]\phi(x)\psi(y) \tag{B-76c}$$

$$= \frac{\sigma_1^2}{2}\left[\sum_{x\in\mathcal{X}}\frac{\phi'(x)}{\sqrt{P_X(x)}}\cdot\phi^2(x) - \sigma_1\sum_{y\in\mathcal{Y}}\frac{\psi'(y)}{\sqrt{P_Y(y)}}\cdot\psi^2(y)\right] \tag{B-76d}$$

$$= \frac{\sigma_1^2}{2}\left[\frac{1}{\sqrt{P_X(1)}}\sum_{x\in\mathcal{X}}\phi'(x)\phi^2(x) - \frac{\operatorname{sgn}(p_1-p_2)\cdot\sigma_1}{\sqrt{P_Y(1)}}\sum_{y\in[d]}\phi'(y)\phi^2(y)\right] \tag{B-76e}$$

$$= \frac{\sigma_1^2}{2}\left[\frac{1}{\sqrt{P_X(1)}} - \frac{\operatorname{sgn}(p_1 - p_2)\cdot\sigma_1}{\sqrt{P_Y(1)}}\right]\sum_{x\in\mathcal{X}}\phi'(x)\phi^2(x). \tag{B-76f}$$

其中 $\boldsymbol{\psi} \triangleq \sigma_1^{-1}\tilde{\mathbf{B}}\boldsymbol{\phi}$，且 (B-76c) 成立基于 $\langle\boldsymbol{\phi},\boldsymbol{\phi}_d\rangle = 0$，(B-76d) 成立基于 $\tilde{\mathbf{B}}\boldsymbol{\phi} = \sigma_1\boldsymbol{\psi}$ 以及 $\tilde{\mathbf{B}}^{\mathsf{T}}\boldsymbol{\psi} = \sigma_1\boldsymbol{\phi}$，(B-76e) 的结果基于引理 B.3 以及

$$P_X(1) = P_X(2) = \cdots = P_X(d),$$
$$P_Y(1) = P_Y(2) = \cdots = P_Y(d).$$

进一步地，为最大化 (B-76f)，注意到

$$\begin{aligned}\sum_{x\in\mathcal{X}}\phi'(x)\phi^2(x) &= \sqrt{\frac{d-1}{d}}\left[\phi^2(1) - \frac{1}{d-1}\sum_{i=2}^{d}\phi^2(i)\right]\\ &= \sqrt{\frac{d-1}{d}}\left[\phi^2(1) - \frac{\|\boldsymbol{\phi}\|^2 - \phi^2(1)}{d-1}\right]\\ &= \sqrt{\frac{d}{d-1}}\phi^2(1) - \frac{1}{\sqrt{(d-1)d}}.\end{aligned}$$

因此由引理 B.4 知当 $\boldsymbol{\phi} = \pm\boldsymbol{\phi}'$ 时 (B-76f) 取得最大值，即 $\boldsymbol{\varphi}_1 = \pm\boldsymbol{\phi}'$，证毕。 □

B.9　推广的 ACE 算法 (4-31)

首先，对 $i = 1,\cdots,k$，我们定义相应的 $|\mathcal{X}|$ 维及 $|\mathcal{Y}|$ 维信息向量 $\bar{\boldsymbol{\phi}}_i$ 与 $\bar{\boldsymbol{\psi}}_i$ 为 (参考定义 2.1)

$$\bar{\phi}_i(x) = \sqrt{\bar{P}_X(x)}\mathbf{f}_i(x), \quad \bar{\psi}_i(y) = \sqrt{\bar{P}_Y(y)}\mathbf{g}_i(y),$$

其中 $\bar{P}_X$ 及 $\bar{P}_Y$ 为 $\bar{P}_{XY}$ 的边缘分布，$\mathbf{f}_i$ 及 $\mathbf{g}_i$ 分别为 $\mathbf{f}$ 及 $\mathbf{g}$ 的第 i 个分量，即对所有 x, y，有

$$\mathbf{f}(x) = \begin{bmatrix} f_1(x) & \cdots & f_k(x)\end{bmatrix}^{\mathsf{T}}, \quad \mathbf{g}(y) = \begin{bmatrix} g_1(y) & \cdots & g_k(y)\end{bmatrix}^{\mathsf{T}}.$$

于是推广的 ACE 算法 (4-31) 的迭代过程可表为

$$\begin{aligned}&\text{i)} \quad \bar{\boldsymbol{\Phi}}_k \leftarrow \bar{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Psi}}_k\left(\bar{\boldsymbol{\Psi}}_k^{\mathsf{T}}\bar{\boldsymbol{\Psi}}_k\right)^{-1}\\ &\text{ii)} \quad \bar{\boldsymbol{\Psi}}_k \leftarrow \bar{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\left(\bar{\boldsymbol{\Phi}}_k^{\mathsf{T}}\bar{\boldsymbol{\Phi}}_k\right)^{-1}\end{aligned} \tag{B-77}$$

其中

$$\bar{\boldsymbol{\Phi}}_k = \left[\bar{\boldsymbol{\phi}}_1,\cdots,\bar{\boldsymbol{\phi}}_k\right], \quad \bar{\boldsymbol{\Psi}}_k = \left[\bar{\boldsymbol{\psi}}_1,\cdots,\bar{\boldsymbol{\psi}}_k\right].$$

注意到 (B-77) 的形式与求解低秩恢复问题

$$\underset{\bar{\boldsymbol{\Psi}}_k,\bar{\boldsymbol{\Phi}}_k}{\text{minimize}} \ \left\|\bar{\mathbf{B}}-\bar{\boldsymbol{\Psi}}_k\bar{\boldsymbol{\Phi}}_k^{\mathsf{T}}\right\|_{\mathrm{F}}^2.$$

的交替最小二乘 (Alternating Least Squares) 算法[55] 一致。因此由 Eckart–Young–Mirsky 定理[56]，推广的 ACE 算法 (4-31) 实质上是在计算矩阵 $\bar{\mathbf{B}}$ 前 k 个奇异值对应的奇异向量。

B.10　引理 4.5 的证明

对任意 $\bar{P}_{XY}\in\bar{\mathcal{N}}(\epsilon)$，设其对应的经验分布分别为 $\hat{P}_{XY}$ 及 Q_X，则由 (4-39) 有

$$D\big(\hat{P}_{XY}\big\|P_{XY}\big)\leqslant\frac{\epsilon}{\bar{\alpha}_k(r)}\quad\text{以及}\quad D(Q_X\|P_X)\leqslant\frac{\epsilon}{r\bar{\alpha}_k(r)}.$$

在此基础上，与 (B-13) 同理，可得

$$\max_{x\in\mathcal{X},y\in\mathcal{Y}}|\Gamma(y,x)|\leqslant\sqrt{\frac{2}{p_{\min}\bar{\alpha}_k(r)}}\quad\text{以及}\quad\max_{x\in\mathcal{X}}|\zeta(x)|\leqslant\sqrt{\frac{2}{rp_{\min}\bar{\alpha}_k(r)}}.$$

此外，由 (4-42) 可推出

$$\begin{aligned}|\Upsilon(y,x)|&\leqslant|\Gamma(y,x)|+\frac{r}{1+r}|\zeta(x)|+|\mathcal{Y}|\cdot\max_{y'\in\mathcal{Y}}|\Gamma(y',x)|\\&\leqslant(|\mathcal{Y}|+1)\cdot\max_{x\in\mathcal{X},y\in\mathcal{Y}}|\Gamma(y,x)|+\sqrt{r}\cdot\max_{x\in\mathcal{X}}|\zeta(x)|\\&\leqslant(|\mathcal{Y}|+2)\sqrt{\frac{2}{p_{\min}\bar{\alpha}_k(r)}},\end{aligned}\tag{B-78}$$

其中第二个不等号由 $\frac{r}{1+r}\leqslant\frac{\sqrt{r}}{2}\leqslant\sqrt{r}$ 推出。

于是由 (B-18) 得

$$|\bar{\Xi}(y,x)|\leqslant\frac{1+|\mathcal{X}|+|\mathcal{Y}|}{p_{\min}^2}\cdot\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|\Upsilon(y,x)|\tag{B-79}$$

$$\leqslant\frac{1+|\mathcal{X}|+|\mathcal{Y}|}{p_{\min}^2}\cdot(|\mathcal{Y}|+2)\cdot\sqrt{\frac{2}{p_{\min}\bar{\alpha}_k(r)}}\tag{B-80}$$

$$\leqslant\frac{(1+|\mathcal{X}|+|\mathcal{Y}|)^2}{p_{\min}^3}\sqrt{\frac{2}{\bar{\alpha}_k(r)}}.\tag{B-81}$$

从而得 $\|\bar{\boldsymbol{\Xi}}\|_{\mathrm{F}}\leqslant\bar{C}$，其中 $\bar{C}\triangleq\frac{(1+|\mathcal{X}|+|\mathcal{Y}|)^2}{p_{\min}^3}\sqrt{\frac{2|\mathcal{X}||\mathcal{Y}|}{\bar{\alpha}_k(r)}}$。

现在开始证明引理的第二部分。为方便表述，以 τ 表示 $\sqrt{\epsilon}$，则由 (4-30)、(4-40) 及 (B-20) 可得出

$$\bar{P}_X(x) = \frac{1}{r+1}\hat{P}_X(x) + \frac{r}{r+1}Q_X(x) = P_X(x) + \tau\sqrt{P_X(x)} \cdot \frac{\Gamma_X(x) + r\zeta(x)}{1+r},$$

其中 Γ_X 与 ζ 分别由 (B-20) 及 (4-40) 给出。此外，由 (4-15) 及 (B-20) 可得

$$\begin{aligned}\hat{P}_{Y|X}(y|x) &= \frac{\hat{P}_{XY}(x,y)}{\hat{P}_X(x)} \\ &= P_{Y|X}(y|x) + \tau\sqrt{P_{Y|X}(y|x)} \cdot \left[\frac{\Gamma(y,x)}{\sqrt{P_X(x)}} - \sqrt{P_{XY}(x,y)} \cdot \frac{\Gamma_X(x)}{P_X(x)}\right] + o(\tau),\end{aligned}$$

其中 Γ 定义由 (4-15) 给出。

故可得

$$\begin{aligned}\bar{P}_{XY}(x,y) &= \bar{P}_X(x)\hat{P}_{Y|X}(y|x) \\ &= P_{XY}(x,y) + \tau\sqrt{P_{XY}(x,y)} \cdot \Big(\Gamma(y,x) \\ &\qquad + \frac{r}{1+r}\sqrt{P_{Y|X}(y|x)} \cdot \big[\zeta(x) - \Gamma_X(x)\big]\Big) + o(\tau) \\ &= P_{XY}(x,y) + \tau\sqrt{P_{XY}(x,y)}\Upsilon(y,x) + o(\tau).\end{aligned}$$

最后，根据 (B-20)–(B-27) 得

$$\bar{\mathbf{B}} = \tilde{\mathbf{B}} + \tau\bar{\mathbf{\Xi}} + o(\tau) = \tilde{\mathbf{B}} + \sqrt{\epsilon}\,\bar{\mathbf{\Xi}} + o\left(\sqrt{\epsilon}\right).$$

B.11　引理 4.6 的证明

我们有

$$\begin{aligned}\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\mathbf{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\mathbf{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} &= \sum_{T(\hat{P}_{XY}),\, T(Q_X):\ \bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)} \mathbb{P}_{n,m}\big\{T(\hat{P}_{XY}), T(Q_X)\big\} \\ &= \sum_{T(\hat{P}_{XY}),\, T(Q_X):\ \bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)} \mathbb{P}_n\big\{T(\hat{P}_{XY})\big\}\mathbb{P}_m\left\{T(Q_X)\right\},\end{aligned} \tag{B-82}$$

其中 $T(\hat{P}_{XY})$ 及 $T(Q_X)$ 分别表示 $\hat{P}_{XY}$ 与 Q_X 所对应的型类，且最后的等式基于 Q_X 与 $\hat{P}_{XY}$ 互相独立。这两个型类的概率分别为[9]

$$\mathbb{P}_n\big\{T(\hat{P}_{XY})\big\} \doteq \exp\left\{-nD(\hat{P}_{XY}\|P_{XY})\right\}$$

以及

$$\mathbb{P}_m\{T(Q_X)\} \doteq \exp\left\{-mD(Q_X\|P_X)\right\} = \exp\left\{-nrD(Q_X\|P_X)\right\}.$$

此外，注意到型的数量至多以 n 的多项式规模增长，因此由 Laplace 原理[92] 有

$$\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} \doteq \exp\Big\{-n\cdot \inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)}\Big[D(\hat{P}_{XY}\|P_{XY}) + rD(Q_X\|P_X)\Big]\Big\}.$$

B.12　引理 4.7 的证明

对于给定 $\epsilon > 0$ 即 $t > 0$, 定义 $\bar{\mathcal{N}}(\epsilon)$ 的子集 $\bar{\mathcal{S}}_2^{(t)}(\epsilon)$ 为

$$\bar{\mathcal{S}}_2^{(t)}(\epsilon) \triangleq \left\{\bar{P}_{XY}:\ \bar{P}_{XY}\in\bar{\mathcal{N}}(\epsilon),\ \sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2} \geqslant t\right\}, \tag{B-83}$$

其中 $\bar{\boldsymbol{\Xi}}$ 定义由 (4-41) 给出。关于该集合，我们有如下引理：

引理 B.5： 对任意 $t\in(0,2)$，有

$$-\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_2^{(t)}(\epsilon)}\left[D\big(\hat{P}_{XY}\big\|P_{XY}\big) + rD(Q_X\|P_X)\right] = \frac{t}{2\bar{\alpha}_k(r)}. \tag{B-84}$$

基于引理 B.5，引理 4.7 可证明如下。首先，对任意 $\bar{P}_{XY}\in\bar{\mathcal{N}}(\epsilon)$，由 引理 4.5 可得

$$\bar{\mathbf{B}}^{\mathsf{T}}\bar{\mathbf{B}} = \tilde{\mathbf{B}}^{\mathsf{T}}\tilde{\mathbf{B}} + \epsilon\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right) + o\left(\sqrt{\epsilon}\right).$$

根据引理 4.1 中的微扰分析结果，学习误差可表为

$$\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 = \epsilon\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2} + o(\epsilon). \tag{B-85}$$

因此任取 $t\in(0,1)$，存在 $\epsilon_0 > 0$ 使得对任意 $\epsilon\in(0,\epsilon_0)$，有

$$\bar{\mathcal{S}}_2^{(1+t)}(\epsilon) \subseteq \bar{\mathcal{S}}_1(\epsilon)\cap\bar{\mathcal{N}}(\epsilon) \subseteq \bar{\mathcal{S}}_2^{(1-t)}(\epsilon).$$

于是基于类似 (B-33)–(B-35) 的推导过程，由引理 B.5 可得

$$\begin{aligned}
&-\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)\cap\bar{\mathcal{N}}(\epsilon)}\left[D\big(\hat{P}_{XY}\big\|P_{XY}\big) + rD(Q_X\|P_X)\right]\\
&= -\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_2^{(1)}(\epsilon)}\left[D\big(\hat{P}_{XY}\big\|P_{XY}\big) + rD(Q_X\|P_X)\right] = \frac{1}{2\bar{\alpha}_k(r)}.
\end{aligned}$$

接下来只需证明引理 B.5。

证明 (引理 B.5 的证明) 因为 $\bar{\mathcal{S}}_2^{(t)}(\epsilon)$ 为闭集，可得

$$\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_2^{(t)}(\epsilon)}\left[D\big(\hat{P}_{XY}\big\|P_{XY}\big)+rD(Q_X\|P_X)\right]=\min_{\bar{P}_{XY}\in\bar{\mathcal{S}}_2^{(t)}(\epsilon)}\left[D\big(\hat{P}_{XY}\big\|P_{XY}\big)+rD(Q_X\|P_X)\right].$$

接着，对任意 $\bar{P}_{XY}\in\bar{\mathcal{S}}_2^{(t)}$ 及相应的有标签数据经验分布 $\hat{P}_{XY}\leftrightarrow\Gamma$ 和无标签数据经验分布 $Q_X\leftrightarrow\zeta$，由 K-L 散度的二阶 Taylor 级数展开有

$$D(\hat{P}_{XY}\|P_{XY})+rD(Q_X\|P_X)=\frac{\epsilon}{2}\left[\|\boldsymbol{\Gamma}\|_{\mathsf{F}}^2+r\|\boldsymbol{\zeta}\|^2\right]+o(\epsilon),\tag{B-86}$$

因此，误差指数 (4-35) 可通过如下优化问题求解：

$$\underset{\boldsymbol{\Gamma},\boldsymbol{\zeta}}{\text{minimize}}\quad \|\boldsymbol{\Gamma}\|_{\mathsf{F}}^2+r\|\boldsymbol{\zeta}\|^2\tag{B-87a}$$

$$\text{subject to}\quad \sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\geqslant t,\tag{B-87b}$$

$$\sum_{x\in\mathcal{X}}\sqrt{P_X(x)}\zeta(x)=0,\tag{B-87c}$$

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\sqrt{P_{XY}(x,y)}\Gamma(y,x)=0,\tag{B-87d}$$

其中等式约束源于 Γ 及 ζ 的定义。尽管优化问题中未考虑约束条件 $\bar{P}_{XY}\in\bar{\mathcal{N}}(\epsilon)$，下面将验证最优的 $(\boldsymbol{\Gamma},\boldsymbol{\zeta})$ 可使得该条件自动满足。由于该优化问题中目标函数及不等式约束条件 (B-87) 均为二次的，其最优解可通过求解如下优化问题求得：

$$\underset{\boldsymbol{\Gamma},\boldsymbol{\zeta}}{\text{maximize}}\quad \sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\tag{B-88a}$$

$$\text{subject to}\quad \|\boldsymbol{\Gamma}\|_{\mathsf{F}}^2+r\|\boldsymbol{\zeta}\|^2\leqslant 1,\tag{B-88b}$$

$$\sum_{x\in\mathcal{X}}\sqrt{P_X(x)}\zeta(x)=0,\tag{B-88c}$$

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\sqrt{P_{XY}(x,y)}\Gamma(y,x)=0,\tag{B-88d}$$

其中我们再次运用了对调目标函数及不等式约束中的二次函数的方法。接着，使用与有监督情形时类似的论证方法，可验证 (B-88) 的最优解也满足 (4-15) 及 (4-42)。在此基础上，可知 (B-88) 等价于除去等式约束后的优化问题，即

$$\underset{\boldsymbol{\Gamma},\boldsymbol{\zeta}}{\text{maximize}}\quad \sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\tag{B-89a}$$

$$\text{subject to} \quad \|\boldsymbol{\Gamma}\|_{\mathsf{F}}^2 + r\|\boldsymbol{\zeta}\|^2 \leqslant 1. \tag{B-89b}$$

实际上，可令 $(\boldsymbol{\Gamma}^*, \boldsymbol{\zeta}^*)$ 为 (B-89) 的最优解并定义 $c_1 \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Gamma^*(y,x)\sqrt{P_{XY}(x,y)}$ 及 $c_2 \triangleq \sum_{x\in\mathcal{X}} \zeta^*(x)$。在此基础上，令 $z_1(x,y) \triangleq \Gamma^*(y,x) - c_1\sqrt{P_{XY}(x,y)}$ 及 $z_2(x,y) \triangleq \zeta^*(x) - c_2\sqrt{P_X(x)}$，则有

$$1 = \|\boldsymbol{\Gamma}^*\|_{\mathsf{F}}^2 + r\|\boldsymbol{\zeta}^*\|^2 = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} z_1^2(x,y) + r\sum_{x\in\mathcal{X}} z_2^2(x) + (c_1^2 + rc_2^2), \tag{B-90}$$

由此可得 $c_1^2 + rc_2^2 \leqslant 1$.

下面对 $c_1^2 + rc_2^2$ 取值进行讨论。若 $c_1^2 + rc_2^2 = 1$，可知 $z_1(x,y) \equiv 0$ 以及 $z_2(x) \equiv 0$，从而得 $\Gamma^*(y,x) = c_1\sqrt{P_{XY}(x,y)}$ 及 $\zeta^*(x) = c_2\sqrt{P_X(x)}$。因此，由 (4-41) 推出

$$\bar{\Xi}(y,x) = -\frac{c_1 + c_2 r}{1+r}\sqrt{P_X(x)P_Y(y)}.$$

从而可知 $\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}$ 为零矩阵。故 (B-88) 目标函数值为零，与 $(\boldsymbol{\Gamma}^*, \boldsymbol{\zeta}^*)$ 最优的假设矛盾。此外，若 $c_1^2 + rc_2^2 < 1$，则可构造可行解 $(\boldsymbol{\Gamma}', \boldsymbol{\zeta}')$ 使得

$$\Gamma'(y,x) = \frac{z_1(x,y)}{\sqrt{1 - c_1^2 - rc_2^2}} \quad \text{及} \quad \zeta'(x) = \frac{z_2(x)}{\sqrt{1 - c_1^2 - rc_2^2}}.$$

易验证 $(\boldsymbol{\Gamma}', \boldsymbol{\zeta}')$ 对应的目标函数值为 $(\boldsymbol{\Gamma}^*, \boldsymbol{\zeta}^*)$ 所对应值的 $\left(1 - c_1^2 - rc_2^2\right)^{-1}$ 倍，再次与 $(\boldsymbol{\Gamma}^*, \boldsymbol{\zeta}^*)$ 的最优性矛盾。综上，必然有 $c_1 = c_2 = 0$，因此优化问题 (B-89) 与 (B-88) 最优解相同。

为化简优化问题 (B-89)，定义向量 $\boldsymbol{\varsigma} \in \mathbb{R}^{|\mathcal{X}|(|\mathcal{Y}|+1)}$ 为

$$\boldsymbol{\varsigma} \triangleq \begin{bmatrix} \operatorname{vec}(\boldsymbol{\Gamma}) \\ \sqrt{r}\boldsymbol{\zeta} \end{bmatrix}, \tag{B-91}$$

并令 $\boldsymbol{\Upsilon}$ 为对应位置元素为 $\Upsilon(y,x)$ 的 $|\mathcal{Y}| \times |\mathcal{X}|$ 矩阵，则由 (4-37) 及 (4-42) 可得 $\operatorname{vec}(\boldsymbol{\Upsilon}) = \bar{\mathbf{L}}(r)\boldsymbol{\varsigma}$。

故 (B-89) 的目标函数可表为

$$\sum_{i=1}^{k}\sum_{j=k+1}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} = \operatorname{vec}^{\mathsf{T}}(\bar{\boldsymbol{\Xi}})\left(\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2 - \sigma_j^2}\right)\operatorname{vec}(\bar{\boldsymbol{\Xi}}), \tag{B-92a}$$

$$= \operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Upsilon})\mathbf{L}^{\mathsf{T}}\left(\sum_{i=1}^{k}\sum_{j=k+1}^{d}\frac{\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{ij}^{\mathsf{T}}}{\sigma_i^2 - \sigma_j^2}\right)\mathbf{L}\operatorname{vec}(\boldsymbol{\Upsilon}) \tag{B-92b}$$

$$= \operatorname{vec}^{\mathsf{T}}(\boldsymbol{\Upsilon})\mathbf{G}_k \operatorname{vec}(\boldsymbol{\Upsilon}) \tag{B-92c}$$

$$= \boldsymbol{\varsigma}^{\mathsf{T}}\bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{G}_k\bar{\mathbf{L}}(r)\boldsymbol{\varsigma} \tag{B-92d}$$

$$= \boldsymbol{\varsigma}^{\mathsf{T}}\bar{\mathbf{G}}_k(r)\boldsymbol{\varsigma}, \tag{B-92e}$$

其中 (B-92a) 基于 (B-41)，且 (B-92c) 基于 (4-10)。此外，根据 $\|\boldsymbol{\varsigma}\|^2 = \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 + r\|\boldsymbol{\zeta}\|^2$, (B-89) 的约束可表为 $\|\boldsymbol{\varsigma}\| \leqslant 1$.

因此 (B-92e) 最大值为 $\bar{\mathbf{G}}_k(r)$ 的谱范数，即 $\bar{\alpha}_k(r)$，同时也是 (B-89) 及 (B-88) 目标函数的最优值。由此可知原优化问题 (B-87) 最优解为 $\left(\sqrt{\frac{t}{\bar{\alpha}_k(r)}}\boldsymbol{\Gamma}^*, \sqrt{\frac{t}{\bar{\alpha}_k(r)}}\boldsymbol{\zeta}^*\right)$，对应最优值为 $t/\bar{\alpha}_k(r)$。令 $\hat{P}_{XY}^* \leftrightarrow \sqrt{\frac{t}{\alpha_k}}\Gamma^*$ 及 $Q_X^* \leftrightarrow \sqrt{\frac{t}{\alpha_k}}\zeta^*$ 为相应的经验分布，则对充分小的 ϵ 有

$$D\big(\hat{P}_{XY}^*\big\|P_{XY}\big) + rD(Q_X^*\|P_X) = \frac{\epsilon t}{2\bar{\alpha}_k(r)} + o(\epsilon) < \frac{\epsilon}{\bar{\alpha}_k(r)},$$

其中最后的不等式基于 $t \in (0,2)$。

因此，相应的由 (4-33) 定义的最优分布 $\bar{P}_{XY}^*$ 满足 $\bar{P}_{XY}^* \in \bar{\mathcal{N}}(\epsilon)$，故

$$\min_{\bar{P}_{XY}\in\bar{\mathcal{S}}_2^{(t)}(\epsilon)} \big[D\big(\hat{P}_{XY}\big\|P_{XY}\big) + rD(Q_X\|P_X)\big] = \frac{\epsilon t}{2\bar{\alpha}_k(r)} + o(\epsilon),$$

由此推出 (B-84)。 □

B.13　定理 4.5的证明

首先由 (4-48) 知存在 $\bar{\epsilon}_1 > 0$ 使得对任意 $\epsilon \in (0, \bar{\epsilon}_1)$ 有

$$\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)}\Big[D(\hat{P}_{XY}\|P_{XY}) + rD(Q_X\|P_X)\Big] = \frac{\epsilon}{2\bar{\alpha}_k(r)} + o(\epsilon) > \frac{\epsilon}{3\bar{\alpha}_k(r)}.$$

于是由 (B-82) 可得

$$\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\}$$

$$= \sum_{T(\hat{P}_{XY}),\, T(Q_X):\ \bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)} \mathbb{P}_n\big\{T(\hat{P}_{XY})\big\}\mathbb{P}_m\left\{T(Q_X)\right\} \tag{B-93a}$$

$$\leqslant \sum_{T(\hat{P}_{XY}),\, T(Q_X):\ \bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)} \exp\left\{-nD(\hat{P}_{XY}\|P_{XY}) - nrD(Q_X\|P_X)\right\} \tag{B-93b}$$

$$\leqslant \sum_{T(\hat{P}_{XY}),\, T(Q_X):\ \bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)} \exp\Big\{-n\cdot\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_1(\epsilon)}\Big[D(\hat{P}_{XY}\|P_{XY}) + rD(Q_X\|P_X)\Big]\Big\} \tag{B-93c}$$

$$\leqslant (n+1)^{|\mathcal{X}||\mathcal{Y}|}(nr+1)^{|\mathcal{X}|}\exp\Big\{-n\cdot\inf_{\bar{P}_{XY}\in\bar{S}_1(\epsilon)}\Big[D(\hat{P}_{XY}\|P_{XY})+rD(Q_X\|P_X)\Big]\Big\} \tag{B-93d}$$

$$< (n+1)^{|\mathcal{X}||\mathcal{Y}|}(nr+1)^{|\mathcal{X}|}\exp\left(-\frac{n\epsilon}{3\bar{\alpha}_k(r)}\right) \tag{B-93e}$$

$$= (n+1)^{|\mathcal{X}||\mathcal{Y}|}(n+r^{-1})^{|\mathcal{X}|}r^{|\mathcal{X}|}\exp\left(-\frac{n\epsilon}{3\bar{\alpha}_k(r)}\right) \tag{B-93f}$$

$$\leqslant (n+\gamma)^{|\mathcal{X}|(|\mathcal{Y}|+1)}r^{|\mathcal{X}|}\exp\left(-\frac{n\epsilon}{3\bar{\alpha}_k(r)}\right), \tag{B-93g}$$

其中 $T(\hat{P}_{XY})$ 与 $T(Q_X)$ 分别表示 $\hat{P}_{XY}$ 与 Q_X 所对应的型类，且 (B-93b) 根据型类概率的上界 (参见 [9] 的定理 11.1.4)，(B-93d) 由型类数量的上界 (参见 [9] 的定理 11.1.1) 给出，且 (B-93g) 中 γ 定义为 $\gamma \triangleq \max\{1, r^{-1}\}$。

从而只需选择 n 使其满足

$$(n+\gamma)^{|\mathcal{X}|(|\mathcal{Y}|+1)}r^{|\mathcal{X}|}\exp\left(-\frac{n\epsilon}{3\bar{\alpha}_k(r)}\right) < \delta,$$

亦即

$$n > \frac{1}{\bar{\mu}}W_{-1}\left(\bar{\mu}e^{\gamma\bar{\mu}}\delta^{\frac{1}{|\mathcal{X}|(|\mathcal{Y}|+1)}}r^{-\frac{1}{|\mathcal{Y}|+1}}\right)-\gamma, \tag{B-94}$$

其中 $\bar{\mu} \triangleq -\frac{\epsilon}{3|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}$，且 $W_{-1}(\cdot)$ 表示下半支的 Lambert W 函数[95]。

此外，定义 $\bar{\epsilon}_2 \triangleq \min\left\{3|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)r^{\frac{1}{|\mathcal{Y}|+1}}|x_0|, \frac{\bar{\alpha}_k(r)\gamma_0}{3|\mathcal{X}|(|\mathcal{Y}|+1)e^{\gamma}}\right\}$，其中 x_0 定义由 (B-46) 给出，且 $\gamma_0 \triangleq \min\{1, r^{-1}\}$。于是对任意 $\epsilon \in (0, \bar{\epsilon}_2)$，有

$$\left|\bar{\mu}e^{\gamma\bar{\mu}}\delta^{\frac{1}{|\mathcal{X}|(|\mathcal{Y}|+1)}}r^{-\frac{1}{|\mathcal{Y}|+1}}\right| < |\bar{\mu}|\,r^{-\frac{1}{|\mathcal{Y}|+1}} = \frac{\epsilon}{3|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}r^{-\frac{1}{|\mathcal{Y}|+1}} < |x_0|,$$

以及

$$\frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon}\log\frac{3|\mathcal{X}|(|\mathcal{Y}|+1)\epsilon}{\bar{\alpha}_k(r)}+\frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon}\log r+\frac{\gamma}{3}$$

$$< \frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon}\log\gamma_0-\frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)\gamma}{\epsilon}+\frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon}\log r+\frac{\gamma}{3} \tag{B-95a}$$

$$\leqslant \frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon}\log\gamma_0+\frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon}\log r-\frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)\gamma}{\epsilon}+\frac{\gamma}{3} \tag{B-95b}$$

$$= \frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon}\log(\gamma_0 r)-\frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)\gamma}{\epsilon}+\frac{\gamma}{3} \tag{B-95c}$$

$$\leqslant \left[\frac{1}{3}-\frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon}\right]\gamma \tag{B-95d}$$

$$\leqslant \left[\frac{1}{3}-\frac{12|\mathcal{X}|^2(|\mathcal{Y}|+1)^2e^{\gamma}}{\gamma_0}\right]\gamma \tag{B-95e}$$

$$< \left[\frac{1}{3} - 12|\mathcal{X}|^2(|\mathcal{Y}|+1)^2 e\right]\gamma \tag{B-95f}$$

$$< 0, \tag{B-95g}$$

其中 (B-95a) 成立基于 $\epsilon < \bar{\epsilon}_2 < \frac{\bar{\alpha}_k(r)\gamma_0}{3|\mathcal{X}|(|\mathcal{Y}|+1)e^{\gamma}}$，(B-95b) 基于 $\gamma_0 \leqslant 1$，(B-95d) 基于 $\gamma_0 r \leqslant 1$，(B-95e) 同样基于 $\epsilon < \frac{\bar{\alpha}_k(r)\gamma_0}{3|\mathcal{X}|(|\mathcal{Y}|+1)e^{\gamma}}$，而 (B-95f) 成立的依据为 $\gamma_0 \leqslant 1 \leqslant \gamma$。

故

$$\frac{1}{\bar{\mu}} W_{-1}\left(\bar{\mu} e^{\gamma\bar{\mu}} \delta^{\frac{1}{|\mathcal{X}|(|\mathcal{Y}|+1)}} r^{-\frac{1}{|\mathcal{Y}|+1}}\right) - \gamma \tag{B-96a}$$

$$< \frac{4}{3\bar{\mu}} \log\left(-\bar{\mu} e^{\gamma\bar{\mu}} \delta^{\frac{1}{|\mathcal{X}|(|\mathcal{Y}|+1)}} r^{-\frac{1}{|\mathcal{Y}|+1}}\right) - \gamma \tag{B-96b}$$

$$= -\frac{4}{3\bar{\mu}} \log\left(-\frac{1}{\bar{\mu}}\right) - \frac{4}{3\bar{\mu}|\mathcal{X}|(|\mathcal{Y}|+1)} \log\frac{1}{\delta} - \frac{4}{3\bar{\mu}(|\mathcal{Y}|+1)} \log r + \frac{\gamma}{3} \tag{B-96c}$$

$$= \frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon} \log\frac{\bar{\alpha}_k(r)}{\epsilon} + \frac{4\bar{\alpha}_k(r)}{\epsilon} \log\frac{1}{\delta} + \frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon} \log[3|\mathcal{X}|(|\mathcal{Y}|+1)] + \frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon} \log r + \frac{\gamma}{3} \tag{B-96d}$$

$$= \frac{8|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon} \log\frac{\bar{\alpha}_k(r)}{\epsilon} + \frac{4\bar{\alpha}_k(r)}{\epsilon} \log\frac{1}{\delta} + \frac{4|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon} \log\frac{3|\mathcal{X}|(|\mathcal{Y}|+1)\epsilon}{\bar{\alpha}_k(r)} + \frac{4|\mathcal{X}|\bar{\alpha}_k(r)}{\epsilon} \log r + \frac{\gamma}{3} \tag{B-96e}$$

$$< \frac{8|\mathcal{X}|(|\mathcal{Y}|+1)\bar{\alpha}_k(r)}{\epsilon} \log\frac{\bar{\alpha}_k(r)}{\epsilon} + \frac{4\bar{\alpha}_k(r)}{\epsilon} \log\frac{1}{\delta} \tag{B-96f}$$

$$= \bar{N}(\epsilon, \delta, r), \tag{B-96g}$$

其中 (B-96b) 成立依据 (B-46)，(B-96f) 成立基于 (B-95)。

最后，令 $\bar{\epsilon}_0 \triangleq \min\{\bar{\epsilon}_1, \bar{\epsilon}_2\}$，则对任意 $\epsilon \in (0, \bar{\epsilon}_0)$ 及 $n \geqslant \bar{N}(\epsilon, \delta, r)$，由 (B-93) 及 (B-96) 可得

$$\mathbb{P}_{n,m}\left\{\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2 > \epsilon\right\} < (n+\gamma)^{|\mathcal{X}|(|\mathcal{Y}|+1)} r^{|\mathcal{X}|} \exp\left(-\frac{n\epsilon}{3\bar{\alpha}_k(r)}\right) < \delta,$$

证毕。

B.14　命题 4.1 的证明

首先，将 (4-37) 中的矩阵 $\bar{\mathbf{L}}(r)$ 写为 $\bar{\mathbf{L}}(r) = \left[\bar{\mathbf{L}}_1(r), \bar{\mathbf{L}}_2(r)\right]$, 其中 $\bar{\mathbf{L}}_1(r)$ 由 $\bar{\mathbf{L}}(r)$ 的前 $(|\mathcal{X}| \cdot |\mathcal{Y}|)$ 列构成，$\bar{\mathbf{L}}_2(r)$ 由 $\bar{\mathbf{L}}$ 剩下的 $|\mathcal{X}|$ 列构成。则根据 $\bar{\mathbf{L}}(r)$ 的定义可得

$$\bar{\mathbf{L}}_1(r) = \mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r}\mathbf{M}\mathbf{M}^{\mathsf{T}} \tag{B-97a}$$

以及

$$\bar{\mathbf{L}}_2(r) = \frac{\sqrt{r}}{1+r}\mathbf{M}, \tag{B-97b}$$

其中 $\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|}$ 为 $\mathbb{R}^{(|\mathcal{X}|\cdot|\mathcal{Y}|)\times(|\mathcal{X}|\cdot|\mathcal{Y}|)}$ 中的单位阵，$\mathbf{M}$ 定义由 (B-54) 给出。

由此可知

$$\begin{aligned}
\bar{\mathbf{L}}(r)\bar{\mathbf{L}}^{\mathsf{T}}(r) &= \bar{\mathbf{L}}_1(r)\bar{\mathbf{L}}_1^{\mathsf{T}}(r) + \bar{\mathbf{L}}_2(r)\bar{\mathbf{L}}_2^{\mathsf{T}}(r) && \text{(B-98a)}\\
&= \mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{2r}{1+r}\mathbf{M}\mathbf{M}^{\mathsf{T}} + \frac{r^2}{(1+r)^2}\mathbf{M}\mathbf{M}^{\mathsf{T}}\mathbf{M}\mathbf{M}^{\mathsf{T}} + \frac{r}{(1+r)^2}\mathbf{M}\mathbf{M}^{\mathsf{T}} && \text{(B-98b)}\\
&= \mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r}\mathbf{M}\mathbf{M}^{\mathsf{T}} && \text{(B-98c)}\\
&= \bar{\mathbf{L}}_1(r), && \text{(B-98d)}
\end{aligned}$$

其中 (B-98c) 成立是因为 $\mathbf{M}^{\mathsf{T}}\mathbf{M}$ 为 $\mathbb{R}^{|\mathcal{X}|}$ 中的单位阵。

在此基础上，有

$$\begin{aligned}
\bar{\alpha}_k(r) &= \left\|\bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{G}_k\bar{\mathbf{L}}(r)\right\|_{\mathrm{s}} = \left\|\left[\mathbf{G}_k^{\frac{1}{2}}\bar{\mathbf{L}}(r)\right]^{\mathsf{T}}\mathbf{G}_k^{\frac{1}{2}}\bar{\mathbf{L}}(r)\right\|_{\mathrm{s}} && \text{(B-99a)}\\
&= \left\|\mathbf{G}_k^{\frac{1}{2}}\bar{\mathbf{L}}(r)\bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{G}_k^{\frac{1}{2}}\right\|_{\mathrm{s}} && \text{(B-99b)}\\
&= \left\|\mathbf{G}_k^{\frac{1}{2}}\left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r}\mathbf{M}\mathbf{M}^{\mathsf{T}}\right)\mathbf{G}_k^{\frac{1}{2}}\right\|_{\mathrm{s}}, && \text{(B-99c)}
\end{aligned}$$

其中 $\mathbf{G}_k^{\frac{1}{2}}$ 定义为满足 $\mathbf{C}^2 = \mathbf{G}_k$ 的半正定矩阵 $\mathbf{C}$，(B-99b) 的导出利用了谱范数 $\|\cdot\|_{\mathrm{s}}$ 的如下性质：对所有矩阵 $\mathbf{A}$，有

$$\left\|\mathbf{A}\mathbf{A}^{\mathsf{T}}\right\|_{\mathrm{s}} = \left\|\mathbf{A}^{\mathsf{T}}\mathbf{A}\right\|_{\mathrm{s}}. \tag{B-100}$$

其次，由 $\mathbf{M}^{\mathsf{T}}\mathbf{M} = \mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|}$ 可得

$$\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r}\mathbf{M}\mathbf{M}^{\mathsf{T}} = [\mathbf{P}(r)]^2, \tag{B-101}$$

其中

$$\mathbf{P}(r) \triangleq \mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \left(1 - \frac{1}{\sqrt{1+r}}\right)\mathbf{M}\mathbf{M}^{\mathsf{T}}.$$

于是可根据 (B-99c) 及 (B-100)–(B-101) 推出

$$\bar{\alpha}_k(r) = \left\|\mathbf{P}(r)\mathbf{G}_k\mathbf{P}(r)\right\|_{\mathrm{s}}.$$

此外，对任意 $r_2 > r_1 \geqslant 0$，定义 $\hat{\mathbf{P}}$ 为

$$\hat{\mathbf{P}} \triangleq \mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \left(1 - \sqrt{\frac{1+r_1}{1+r_2}}\right)\mathbf{M}\mathbf{M}^{\mathsf{T}},$$

则易知 $\hat{\mathbf{P}}$ 满足 $\left\|\hat{\mathbf{P}}\right\|_{\mathrm{s}} = 1$ 及 $\mathbf{P}(r_2) = \mathbf{P}(r_1)\hat{\mathbf{P}} = \hat{\mathbf{P}}\mathbf{P}(r_1)$。因此，

$$\begin{aligned}
\bar{\alpha}_k(r_2) &= \left\|\mathbf{P}(r_2)\mathbf{G}_k\mathbf{P}(r_2)\right\|_{\mathrm{s}} \\
&= \left\|\hat{\mathbf{P}}\mathbf{P}(r_1)\mathbf{G}_k\mathbf{P}(r_1)\hat{\mathbf{P}}\right\|_{\mathrm{s}} \\
&\leqslant \left\|\hat{\mathbf{P}}\right\|_{\mathrm{s}}^2 \left\|\mathbf{P}(r_1)\mathbf{G}_k\mathbf{P}(r_1)\right\|_{\mathrm{s}} \\
&= \left\|\mathbf{P}(r_1)\mathbf{G}_k\mathbf{P}(r_1)\right\|_{\mathrm{s}} = \bar{\alpha}_k(r_1),
\end{aligned}$$

其中的不等式根据的是谱范数次可乘性[63]。

欲证 $\bar{\alpha}_k(r)$ 的凸性，首先对 $r \geqslant 0$ 定义函数 $w(r) = \frac{r}{1+r}$。由于 $w(r)$ 为 r 的增函数及凹函数，对所有的 $r_1, r_2 > 0$ 及 $\theta \in (0,1)$ 有

$$w(\theta r_1 + (1-\theta)r_2) \geqslant \theta w(r_1) + (1-\theta)w(r_2),$$

由此可知

$$\theta r_1 + (1-\theta) \geqslant w^{-1}\left(\theta w(r_1) + (1-\theta)w(r_2)\right).$$

因此，我们有

$$\begin{aligned}
\bar{\alpha}_k\left(\theta r_1 + (1-\theta)\right) &\leqslant \bar{\alpha}_k\left(w^{-1}\left(\theta w(r_1) + (1-\theta)w(r_2)\right)\right) \\
&= \left\|\mathbf{G}_k^{\frac{1}{2}}\left[\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \left(\theta w(r_1) + (1-\theta)w(r_2)\right)\mathbf{M}\mathbf{M}^{\mathsf{T}}\right]\mathbf{G}_k^{\frac{1}{2}}\right\|_{\mathrm{s}} \\
&= \left\|\mathbf{G}_k^{\frac{1}{2}}\left[\theta\left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - w(r_1)\mathbf{M}\mathbf{M}^{\mathsf{T}}\right)\right.\right. \\
&\qquad \left.\left. +(1-\theta)\left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - w(r_2)\mathbf{M}\mathbf{M}^{\mathsf{T}}\right)\right]\mathbf{G}_k^{\frac{1}{2}}\right\|_{\mathrm{s}} \\
&\leqslant \theta\left\|\mathbf{G}_k^{\frac{1}{2}}\left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - w(r_1)\mathbf{M}\mathbf{M}^{\mathsf{T}}\right)\mathbf{G}_k^{\frac{1}{2}}\right\|_{\mathrm{s}} \\
&\qquad + (1-\theta)\left\|\mathbf{G}_k^{\frac{1}{2}}\left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - w(r_2)\mathbf{M}\mathbf{M}^{\mathsf{T}}\right)\mathbf{G}_k^{\frac{1}{2}}\right\|_{\mathrm{s}} \\
&= \theta\bar{\alpha}_k(r_1) + (1-\theta)\bar{\alpha}_k(r_2),
\end{aligned}$$

其中第一个等号利用了 $\bar{\alpha}_k(r)$ 非增的性质，第二个等号依据为谱范数的三角不等式。

最后，为导出 (4-49) 中的下界，注意到

$$\bar{\alpha}_k(r) = \left\| \mathbf{G}_k^{\frac{1}{2}} \left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r} \mathbf{M}\mathbf{M}^{\mathsf{T}} \right) \mathbf{G}_k^{\frac{1}{2}} \right\|_{\mathrm{s}} \tag{B-102a}$$

$$\geqslant \left\| \mathbf{G}_k \right\|_{\mathrm{s}} - \frac{r}{1+r} \left\| \mathbf{G}_k^{\frac{1}{2}} \mathbf{M}\mathbf{M}^{\mathsf{T}} \mathbf{G}_k^{\frac{1}{2}} \right\|_{\mathrm{s}} \tag{B-102b}$$

$$= \left\| \mathbf{G}_k \right\|_{\mathrm{s}} - \frac{r}{1+r} \left\| \mathbf{M}^{\mathsf{T}} \mathbf{G}_k \mathbf{M} \right\|_{\mathrm{s}} \tag{B-102c}$$

$$\geqslant \left\| \mathbf{G}_k \right\|_{\mathrm{s}} - \frac{r}{1+r} \left\| \mathbf{G}_k \right\|_{\mathrm{s}} \left\| \mathbf{M} \right\|_{\mathrm{s}}^2 \tag{B-102d}$$

$$= \frac{1}{1+r} \left\| \mathbf{G}_k \right\|_{\mathrm{s}} = \frac{1}{1+r} \bar{\alpha}_k(0), \tag{B-102e}$$

其中 (B-102b) 依据为三角不等式，(B-102c) 由 (B-100) 推出，(B-102d) 利用了谱范数的次可乘性；为导出倒数第二个等式，注意到由于 $\mathbf{M}^{\mathsf{T}}\mathbf{M}$ 为单位阵，我们有 $\left\| \mathbf{M} \right\|_{\mathrm{s}} = \sqrt{\left\| \mathbf{M}^{\mathsf{T}}\mathbf{M} \right\|_{\mathrm{s}}} = 1$。

为推导 (4-49) 的下界，注意到

$$\begin{aligned}
\bar{\alpha}_k(r) &= \left\| \mathbf{G}_k^{\frac{1}{2}} \left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r} \mathbf{M}\mathbf{M}^{\mathsf{T}} \right) \mathbf{G}_k^{\frac{1}{2}} \right\|_{\mathrm{s}} \\
&= \left\| \frac{1}{1+r} \mathbf{G}_k + \frac{r}{1+r} \mathbf{G}_k^{\frac{1}{2}} \left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \mathbf{M}\mathbf{M}^{\mathsf{T}} \right) \mathbf{G}_k^{\frac{1}{2}} \right\|_{\mathrm{s}} \\
&\leqslant \frac{1}{1+r} \left\| \mathbf{G}_k \right\|_{\mathrm{s}} + \frac{r}{1+r} \left\| \mathbf{G}_k^{\frac{1}{2}} \left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \mathbf{M}\mathbf{M}^{\mathsf{T}} \right) \mathbf{G}_k^{\frac{1}{2}} \right\|_{\mathrm{s}} \\
&= \frac{1}{1+r} \bar{\alpha}_k(0) + \frac{1}{1+r} \bar{\alpha}_k(\infty),
\end{aligned}$$

其中我们再次应用了三角不等式。

B.15　推论 4.3 的证明

首先由 (4-36) 得

$$\begin{aligned}
\bar{\mathbf{G}}_k(r) = \bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{G}_k \bar{\mathbf{L}}(r) &= \frac{1}{4} \sum_{i=1}^{d-1} \sigma_i^2 \left(\bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{M} \boldsymbol{\phi}_i \right) \left(\bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{M} \boldsymbol{\phi}_i \right)^{\mathsf{T}} \\
&= \frac{1}{4(1+r)} \sum_{i=1}^{d-1} \sigma_i^2 \left(\hat{\mathbf{M}}(r) \boldsymbol{\phi}_i \right) \left(\hat{\mathbf{M}}(r) \boldsymbol{\phi}_i \right)^{\mathsf{T}},
\end{aligned} \tag{B-103}$$

其中第二个等号依据 (B-55)，且在最后的等式中定义了

$$\hat{\mathbf{M}}(r) \triangleq \sqrt{1+r} \cdot \bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{M}. \tag{B-104}$$

此外，注意到 $\hat{\mathbf{M}}(r)$ 满足

$$\begin{aligned}\hat{\mathbf{M}}^{\mathsf{T}}(r)\hat{\mathbf{M}}(r) &= (1+r)\mathbf{M}^{\mathsf{T}}\bar{\mathbf{L}}(r)\bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{M} \\ &= (1+r)\mathbf{M}^{\mathsf{T}}\left(\mathbf{I}_{|\mathcal{X}|\cdot|\mathcal{Y}|} - \frac{r}{1+r}\mathbf{M}\mathbf{M}^{\mathsf{T}}\right)\mathbf{M} = \mathbf{M}^{\mathsf{T}}\mathbf{M} = \mathbf{I}_d, \end{aligned} \tag{B-105}$$

其中第二个等号依据 (B-98c)。故得 $\left\langle \hat{\mathbf{M}}(r)\boldsymbol{\phi}_i, \hat{\mathbf{M}}(r)\boldsymbol{\phi}_j \right\rangle = \langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle = \delta_{ij}$，因此由 (B-103) 可知 $\bar{\mathbf{G}}_k(r)$ 的非零特征值为

$$\frac{\sigma_i^2}{4(1+r)}, \quad i = 1, \cdots, d-1.$$

因此 $\bar{\mathbf{G}}_k(r)$ 的最大特征值 (即最大奇异值) 为

$$\bar{\alpha}_k(r) = \left\|\bar{\mathbf{G}}_k(r)\right\|_{\mathrm{s}} = \frac{\sigma_1^2}{4(1+r)}.$$

B.16　定理 4.6 的证明

与定理 4.3 的证明类似，我们首先将 $\bar{\mathcal{N}}(\epsilon)$ 的定义拓展到 $\sigma_k = \sigma_{k+1}$ 的情形：

$$\bar{\mathcal{N}}(\epsilon) \triangleq \left\{ \bar{P}_{XY} :\ D(\hat{P}_{XY}\|P_{XY}) + rD(Q_X\|P_X) \leqslant \frac{\epsilon}{\bar{\beta}_k(r)} \right\}, \tag{B-106}$$

并定义 $\bar{S}_3^{(t)}(\epsilon)$ 为 $\bar{P}_{XY}$ 的集合，使其所对应的 Υ [定义参见 (4-42)] 满足

$$\sum_{i=1}^{l-1}\sum_{j=l}^{d} \frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} + \sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathrm{c}}} \frac{\left[\bar{\boldsymbol{\varphi}}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}} + \bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2 - \sigma_j^2} \geqslant 1, \tag{B-107}$$

其中 $\bar{\boldsymbol{\varphi}}_i$ 的定义参见 (4-52)。类似于 $\sigma_k > \sigma_{k+1}$ 情形时引理 B.5 的结果，我们有如下引理。

引理 B.6： 对任意 $t \in (0,2)$，有

$$-\lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\inf_{\bar{P}_{XY}\in\bar{S}_3^{(t)}(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right) + rD(Q_X\|P_X)\right] = \frac{t}{2\bar{\beta}_k(r)}. \tag{B-108}$$

证明 该引理证明过程与引理 B.2 类似。基于 K-L 散度的二阶 Taylor 展开 (B-86)，极限 (B-108) 的值可通过如下优化问题求解：

$$\underset{\boldsymbol{\Gamma},\boldsymbol{\zeta}}{\text{minimize}} \quad \|\boldsymbol{\Gamma}\|_{\mathsf{F}}^2 + r\|\boldsymbol{\zeta}\|^2 \tag{B-109a}$$

$$\text{subject to}\quad \sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}+\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left[\bar{\boldsymbol{\varphi}}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\geqslant t, \tag{B-109b}$$

$$\sum_{x\in\mathcal{X}}\sqrt{P_X(x)}\zeta(x)=0,\qquad \sum_{x\in\mathcal{X},y\in\mathcal{Y}}\sqrt{P_{XY}(x,y)}\Gamma(y,x)=0. \tag{B-109c}$$

与 引理 B.5 证明同理，(B-109) 的最优解可通过求解如下优化问题求得：

$$\underset{\boldsymbol{\Gamma},\boldsymbol{\zeta}}{\text{maximize}}\quad \sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}+\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left[\bar{\boldsymbol{\varphi}}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2} \tag{B-110a}$$

$$\text{subject to}\quad \|\boldsymbol{\Gamma}\|_{\mathsf{F}}^2+r\|\boldsymbol{\zeta}\|^2\leqslant 1, \tag{B-110b}$$

其中我们对调了目标函数及不等式约束中的二次函数，并去掉了等式约束。

此外，与 (B-92) 类似，(B-110) 的目标函数可表为

$$\sum_{i=1}^{l-1}\sum_{j=l}^{d}\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}+\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left[\bar{\boldsymbol{\varphi}}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}=\boldsymbol{\varsigma}^{\mathsf{T}}\bar{\mathbf{J}}_k(r,\boldsymbol{\Gamma},\boldsymbol{\zeta})\boldsymbol{\varsigma},$$

其中 $\bar{\mathbf{J}}_k(r,\boldsymbol{\Gamma},\boldsymbol{\zeta})$ 定义由 (4-51) 给出。故优化问题 (B-110) 可等价地写为 (4-53)，从而其最优值为 $\bar{\beta}_k(r)$。最后，类似于引理 B.5 的证明过程，可知 (4-53) 最优值为 $t/\bar{\beta}_k(r)$ 从而

$$\inf_{\bar{P}_{XY}\in\bar{\mathcal{S}}_3^{(t)}(\epsilon)}\left[D\left(\hat{P}_{XY}\middle\|P_{XY}\right)+rD(Q_X\|P_X)\right]=\frac{\epsilon t}{2\bar{\beta}_k(r)}+o(\epsilon),$$

由此推出 (B-108)。 □

此外，由引理 4.2 及引理 4.5 可知 $\bar{P}_{XY}\in\mathcal{N}(\epsilon)$ 所对应的学习误差为

$$\begin{aligned}\left\|\tilde{\mathbf{B}}\boldsymbol{\Phi}_k\right\|_{\mathrm{F}}^2-\left\|\tilde{\mathbf{B}}\bar{\boldsymbol{\Phi}}_k\right\|_{\mathrm{F}}^2=\epsilon\sum_{i=1}^{l-1}\sum_{j=l}^{d}&\frac{\left[\boldsymbol{\phi}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}\\&+\epsilon\sum_{i=l}^{k}\sum_{j\in\mathcal{I}_k^{\mathsf{c}}}\frac{\left[\bar{\boldsymbol{\varphi}}_i^{\mathsf{T}}\left(\tilde{\mathbf{B}}^{\mathsf{T}}\bar{\boldsymbol{\Xi}}+\bar{\boldsymbol{\Xi}}^{\mathsf{T}}\tilde{\mathbf{B}}\right)\boldsymbol{\phi}_j\right]^2}{\sigma_i^2-\sigma_j^2}+o(\epsilon).\end{aligned} \tag{B-111}$$

故对任意 $t \in (0,1)$，存在 $\epsilon_0 > 0$ 使得对所有的 $\epsilon \in (0, \epsilon_0)$ 有

$$\bar{\mathcal{S}}_3^{(1+t)}(\epsilon) \subseteq \bar{\mathcal{S}}_1(\epsilon) \cap \bar{\mathcal{N}}(\epsilon) \subseteq \bar{\mathcal{S}}_3^{(1-t)}(\epsilon).$$

于是，与 (B-33)–(B-35) 的推导类似，由引理 B.6 可得

$$\begin{aligned} &- \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \inf_{\bar{P}_{XY} \in \bar{\mathcal{S}}_1(\epsilon) \cap \bar{\mathcal{N}}(\epsilon)} \left[D\big(\hat{P}_{XY} \big\| P_{XY}\big) + rD(Q_X \| P_X) \right] \\ &\quad = - \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \inf_{\bar{P}_{XY} \in \bar{\mathcal{S}}_3^{(1)}(\epsilon)} \left[D\big(\hat{P}_{XY} \big\| P_{XY}\big) + rD(Q_X \| P_X) \right] = \frac{1}{2\bar{\beta}_k(r)}. \end{aligned}$$

最后，与定理 4.4 的证明过程同理，可得 (4-55)。

B.17　推论 4.4的证明

由引理 B.3 可得 $\sigma_1 = \cdots = \sigma_{d-1} > \sigma_d = 0$，故对任意 $1 \leqslant k \leqslant d-1$ 有 $\mathcal{I}_k = [d-1]$，从而

$$l = \min \mathcal{I}_k = 1, \quad \mathcal{I}_k^{\mathsf{c}} = \{d\}.$$

因此，由 (4-51) 可得

$$\bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta}) = \bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{L}^{\mathsf{T}} \left(\sum_{i=1}^{k} \frac{\bar{\boldsymbol{\vartheta}}_{id} \bar{\boldsymbol{\vartheta}}_{id}^{\mathsf{T}}}{\sigma_1^2} \right) \mathbf{L} \bar{\mathbf{L}}(r) = \frac{1}{\sigma_1^2} \sum_{i=1}^{k} \left(\bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{L}^{\mathsf{T}} \bar{\boldsymbol{\vartheta}}_{id} \right) \left(\bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{L}^{\mathsf{T}} \bar{\boldsymbol{\vartheta}}_{id} \right)^{\mathsf{T}}. \tag{B-112}$$

此外，类似于 (B-51)，我们有

$$\mathbf{L}^{\mathsf{T}} \bar{\boldsymbol{\vartheta}}_{id} = -\frac{\sigma_1^2}{2} \mathbf{M} \bar{\boldsymbol{\varphi}}_i, \tag{B-113}$$

由此推出

$$\bar{\mathbf{G}}_k = \frac{\sigma_1^2}{4} \sum_{i=1}^{k} \left(\bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{M} \bar{\boldsymbol{\varphi}}_i \right) \left(\bar{\mathbf{L}}^{\mathsf{T}}(r) \mathbf{M} \bar{\boldsymbol{\varphi}}_i \right)^{\mathsf{T}} = \frac{1}{4(1+r)} \sum_{i=1}^{d-1} \sigma_i^2 \left(\hat{\mathbf{M}}(r) \bar{\boldsymbol{\varphi}}_i \right) \left(\hat{\mathbf{M}}(r) \bar{\boldsymbol{\varphi}}_i \right)^{\mathsf{T}}, \tag{B-114}$$

其中 $\hat{\mathbf{M}}(r)$ 定义由 (B-104) 给出。注意到由于 $\left\langle \hat{\mathbf{M}}(r) \bar{\boldsymbol{\varphi}}_i, \hat{\mathbf{M}}(r) \bar{\boldsymbol{\varphi}}_j \right\rangle = \delta_{ij}$，(B-114) 给出了矩阵 $\bar{\mathbf{G}}_k$ 的特征值分解。因此，结合 定理 4.6 及 $\bar{\beta}(r)$ 的定义，有

$$\bar{\beta}_k(r) \leqslant \left\| \bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta}) \right\|_{\mathrm{s}} = \frac{\sigma_1^4}{4(1+r)}.$$

为证明上式中不等号取得等号，只需构造 $\boldsymbol{\Gamma}$ 及 $\boldsymbol{\zeta}$ 使得相应的由 (4-54) 定义的 $\boldsymbol{\varsigma}$ 满足 $\|\boldsymbol{\varsigma}\|^2 \leqslant 1$ 与

$$\boldsymbol{\varsigma}^{\mathsf{T}} \bar{\mathbf{G}}_k \boldsymbol{\varsigma} = \left\| \bar{\mathbf{J}}_k(r, \boldsymbol{\Gamma}, \boldsymbol{\zeta}) \right\|_{\mathrm{s}}. \tag{B-115}$$

事实上，易知若 $\boldsymbol{\Gamma}$ 及 $\boldsymbol{\zeta}$ 分别取

$$\Gamma(y, x) = \frac{1}{\sqrt{1+r}} \sqrt{P_{Y|X}(y|x)} \phi'(x) \tag{B-116a}$$

与

$$\zeta(x) = \frac{1}{\sqrt{1+r}} \phi'(x), \tag{B-116b}$$

其中 $\boldsymbol{\phi}'$ 的定义参见 (B-67)，则有 $\bar{\boldsymbol{\varphi}}_1 = \pm \boldsymbol{\phi}'$ 以及 $\boldsymbol{\varsigma} = \hat{\mathbf{M}}(r)\boldsymbol{\phi}' = \pm \hat{\mathbf{M}}(r)\bar{\boldsymbol{\varphi}}_1$，因此 (B-115) 成立。

为验证该结论，首先注意到由 (B-97) 有

$$\hat{\mathbf{M}}(r) = \sqrt{1+r}\bar{\mathbf{L}}^{\mathsf{T}}(r)\mathbf{M} = \sqrt{1+r} \begin{bmatrix} \bar{\mathbf{L}}_1^{\mathsf{T}}(r)\mathbf{M} \\ \bar{\mathbf{L}}_2^{\mathsf{T}}(r)\mathbf{M} \end{bmatrix} = \frac{1}{\sqrt{1+r}} \begin{bmatrix} \mathbf{M} \\ \sqrt{r}\mathbf{I}_d \end{bmatrix},$$

于是由 (4-54) 与 (B-116) 中可得 $\boldsymbol{\varsigma} = \hat{\mathbf{M}}(r)\boldsymbol{\phi}'$，从而有 $\|\boldsymbol{\varsigma}\|^2 = \|\boldsymbol{\phi}'\|^2 = 1$。

此外，由 (4-42) 可得

$$\begin{aligned} \Upsilon(y, x) &= \Gamma(y, x) + \frac{r}{1+r} \sqrt{P_{Y|X}(y|x)} \cdot \left[\zeta(x) - \sum_{y'} \sqrt{P_{Y|X}(y'|x)} \Gamma(y', x) \right] \\ &= \Gamma(y, x) = \frac{1}{\sqrt{1+r}} \sqrt{P_{Y|X}(y|x)} \phi'(x), \end{aligned}$$

即

$$\mathrm{vec}(\boldsymbol{\Upsilon}) = \frac{1}{\sqrt{1+r}} \mathbf{M}\boldsymbol{\phi}'.$$

从而类似于 (B-74)，由 (4-41) 可推出

$$\bar{\Xi}(y, x) = \frac{1}{\sqrt{1+r}} \Xi(y, x)$$

其中 $\Xi(y, x)$ 定义如 (B-74) 所示。在此基础上，与 推论 4.2 的证明过程同理，可知 $\bar{\boldsymbol{\varphi}}_1$ 为如下优化问题的解：

$$\underset{\boldsymbol{\phi}}{\text{maximize}} \quad \boldsymbol{\phi}^{\mathsf{T}} \left(\tilde{\mathbf{B}}^{\mathsf{T}} \bar{\boldsymbol{\Xi}} \right) \boldsymbol{\phi}$$

$$\text{subject to} \quad \langle \boldsymbol{\phi}, \boldsymbol{\phi}_d \rangle = 0, \quad \|\boldsymbol{\phi}\| = 1, \tag{B-117}$$

由于 $\bar{\boldsymbol{\Xi}} = \boldsymbol{\Xi}/\sqrt{1+r}$，其解也与 (B-75) 一致。因此得出 $\bar{\boldsymbol{\varphi}}_1 = \pm\boldsymbol{\phi}$’，证毕。

附录 C　第 5 章中的证明

C.1　引理 5.1 的证明

由矩阵行列式引理，可求得矩阵 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 的特征多项式为

$$\begin{aligned} p(\lambda) &= \det(\lambda\mathbf{I}-\mathbf{D}-\mathbf{u}\mathbf{v}^{\mathsf{T}}) \\ &= \det(\lambda\mathbf{I}-\mathbf{D}) - \mathbf{v}^{\mathsf{T}}\operatorname{adj}(\lambda\mathbf{I}-\mathbf{D})\mathbf{u} \\ &= \prod_{i=1}^{m}(\lambda-D_{ii}) - \sum_{i=1}^{m} u_i v_i \prod_{j\neq i}(\lambda-D_{jj}) \\ &= \left(1-\sum_{i=1}^{m}\frac{u_i v_i}{\lambda-D_{ii}}\right)\cdot\prod_{i=1}^{m}(\lambda-D_{ii}), \end{aligned}$$

则 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 的所有特征值由 $p(\lambda)=0$ 的根给出。

首先，证明 $\mathcal{S}_1\cup\mathcal{S}_2\subset\mathcal{S}$，其中 $\mathcal{S}$ 为矩阵 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 所有特征值所构成集合。若 $\lambda\in\mathcal{S}_1$，则显然 $p(\lambda)=0$；若 $\lambda\in\mathcal{S}_2$，则存在 i 使得 $\lambda=D_{ii}$。若 $\exists\, i\neq j$ 满足 $D_{ii}=D_{jj}$，则有 $(\lambda-D_{ii})\mid p(\lambda)$ 从而 $p(D_{ii})=0$；否则有 $u_i v_i=0$，从而

$$p(D_{ii}) = \left(1-\sum_{j\neq i}\frac{u_j v_j}{D_{ii}-D_{jj}}\right)\cdot\prod_{j=1}^{m}(D_{ii}-D_{jj}) = 0,$$

故结论成立。

其次证明 $\mathcal{S}\subset\mathcal{S}_1\cup\mathcal{S}_2$。为此，注意到对任意 $\lambda\in\mathcal{S}$，若 $\lambda\notin\{D_{ii}:i\in[m]\}$ 则 $\prod_{i=1}^{m}(\lambda-D_{ii})\neq 0$ 以及 $\lambda\in\mathcal{S}_1$；否则有 $\lambda=D_{ii}$，其中在所有 D_{ii} 的取值中，若 D_{ii} 取值与其它元素重复，则根据定义有 $\lambda\in\mathcal{S}_2$；否则

$$0 = p(D_{ii}) = -u_i v_i\cdot\prod_{j\neq i}(D_{ii}-D_{jj}),$$

从而得出 $u_i v_i=0$ 以及 $\lambda\in\mathcal{S}_2$。

由上述推导可知，若 $\lambda\in\mathcal{S}_1\setminus\mathcal{S}_2$，则 $\lambda\notin\{D_{ii}:i\in[m]\}$。故所有 $\lambda\in\mathcal{S}_1\setminus\mathcal{S}_2$ 均为 $p(\lambda)=0$ 的单根，所对应矩阵 $(\mathbf{D}-\lambda\mathbf{I})$ 非奇异。为求解其所对应的特征向量 $\mathbf{q}$，由 $(\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}-\lambda\mathbf{I})\mathbf{q}=0$ 得 $\mathbf{q}=\langle\mathbf{q},\mathbf{v}\rangle(\lambda\mathbf{I}-\mathbf{D})^{-1}\mathbf{u}\propto(\lambda\mathbf{I}-\mathbf{D})^{-1}\mathbf{u}$。

最后，为证明 $\mathbf{D}+\mathbf{u}\mathbf{v}^{\mathsf{T}}$ 可对角化，考察其所有重特征值的几何重数。为方便表述，对给定 λ，记 $\mathbf{u}_\lambda\triangleq(u_i:i\in\mathcal{I}_\lambda)^{\mathsf{T}}$ 为 $\mathbf{u}$ 中所有下标在 $\mathcal{I}_\lambda$ 中的元素所构成的子向量；类似地，定义 $\mathbf{v}_\lambda\triangleq(v_i:i\in\mathcal{I}_\lambda)^{\mathsf{T}}$。则引理中条件 (ii) 可重述为：对任意

$\lambda \in \mathcal{S}_2$，有 $\langle \mathbf{u}_\lambda, \mathbf{v}_\lambda \rangle \neq 0$ 或 $\|\mathbf{u}_\lambda\| \cdot \|\mathbf{v}_\lambda\| = 0$ 成立。由于所有的 $\lambda \in \mathcal{S}_1 \setminus \mathcal{S}_2$ 均为单特征值，$\mathbf{D} + \mathbf{u}\mathbf{v}^\mathsf{T}$ 的任一重特征值必定属于集合 $\mathcal{S}_2$。不失一般性，设 $\lambda_1 = D_{11} = D_{22} = \cdots = D_{ll} \notin \{D_{jj} : j > l\}$ 为一代数重数 $\mu(\lambda_1) > 1$ 的重特征值。由 $\lambda_1 \in \mathcal{S}_2$ 及 $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$ 可得 $\lambda_1 \notin \mathcal{S}_1$ 以及

$$p(\lambda) = \left(1 - \sum_{i>l} \frac{u_i v_i}{\lambda - D_{ii}} - \frac{\langle \mathbf{u}_{\lambda_1}, \mathbf{v}_{\lambda_1} \rangle}{\lambda - \lambda_1}\right) \cdot \prod_{i=1}^{m} (\lambda - D_{ii}).$$

从而由条件 (i), (ii) 知存在以下两种可能情况：

1. $\langle \mathbf{u}_{\lambda_1}, \mathbf{v}_{\lambda_1} \rangle \neq 0$;
2. $\langle \mathbf{u}_{\lambda_1}, \mathbf{v}_{\lambda_1} \rangle = 0, 1 - \sum_{i>l} \frac{u_i v_i}{\lambda_1 - D_{ii}} \neq 0$.

易验证第一种情况下 $\mu(\lambda_1) = l - 1$，第二种情况下 $\mu(\lambda_1) = l$。为证明 $\mathbf{D} + \mathbf{u}\mathbf{v}^\mathsf{T}$ 可对角化，只需考察 λ_1 的几何重数 $\gamma(\lambda_1)$ 并证明 $\gamma(\lambda_1) = \mu(\lambda_1)$.

对第一种情况，可得 $\mathbf{u}_{\lambda_1}, \mathbf{v}_{\lambda_1} \neq \mathbf{0}_l$。计算可得 λ_1 所对应的特征空间为 $\{\mathbf{w} = (\hat{\mathbf{w}}^\mathsf{T}, \mathbf{0}_{m-l}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^m : \langle \hat{\mathbf{w}}, \mathbf{v}_{\lambda_1} \rangle = 0\}$，其维度为 $\gamma(\lambda_1) = l - 1 = \mu(\lambda_1)$。

对第二种情况，由假设知 $\mathbf{u}_{\lambda_1} = \mathbf{0}_l$ 或 $\mathbf{v}_{\lambda_1} = \mathbf{0}_l$ 成立。若 $\mathbf{u}_{\lambda_1} \neq \mathbf{0}_l$，可得 $\mathbf{v}_{\lambda_1} = \mathbf{0}_l$ 且 λ_1 的特征空间为 $\left\{\mathbf{w} = (\hat{\mathbf{w}}^\mathsf{T}, \mathbf{0}_{m-l}^\mathsf{T})^\mathsf{T} : \hat{\mathbf{w}} \in \mathbb{R}^l\right\}$，对应几何重数 $\gamma(\lambda_1) = l = \mu(\lambda_1)$；否则，有 $\mathbf{u}_{\lambda_1} = \mathbf{0}_l$ 成立。记 $\hat{\mathbf{D}} \triangleq \operatorname{diag}\{\lambda_1 - D_{l+1,l+1}, \cdots, \lambda_1 - D_{mm}\} \in \mathbb{R}^{(m-l)\times(m-l)}$，$\bar{\mathbf{u}}_{\lambda_1} \triangleq (u_{l+1}, \cdots, u_m)^\mathsf{T} \in \mathbb{R}^{m-l}$，$\bar{\mathbf{v}}_{\lambda_1} \triangleq (v_{l+1}, \cdots, v_m)^\mathsf{T}$，则相应的特征空间可表示为 $\left\{\mathbf{w} = (\hat{\mathbf{w}}^\mathsf{T}, \langle \hat{\mathbf{w}}, \mathbf{v}_{\lambda_1} \rangle \cdot \bar{\mathbf{u}}_{\lambda_1}^\mathsf{T} (\hat{\mathbf{D}} - \bar{\mathbf{u}}_{\lambda_1} \bar{\mathbf{v}}_{\lambda_1}^\mathsf{T})^{-1})^\mathsf{T} : \hat{\mathbf{w}} \in \mathbb{R}^l\right\}$，其维度为 $\gamma(\lambda_1) = l$。注意到 $\hat{\mathbf{D}} - \bar{\mathbf{u}}_{\lambda_1} \bar{\mathbf{v}}_{\lambda_1}^\mathsf{T}$ 非奇异，因

$$\det(\hat{\mathbf{D}} - \bar{\mathbf{u}}_{\lambda_1} \bar{\mathbf{v}}_{\lambda_1}^\mathsf{T}) = \det(\hat{\mathbf{D}}) \cdot \left(1 - \sum_{i>l} \frac{u_i v_i}{\lambda_1 - D_{ii}}\right) \neq 0.$$

C.2　定理 5.3的证明

注意到由 (5-3) 有

$$v(\boldsymbol{\phi}_n) = \frac{\langle \boldsymbol{\phi}_n, \mathbf{v}_1 \rangle^2}{\|\boldsymbol{\phi}_n\|^2} = \frac{\langle \tilde{\boldsymbol{\phi}}_n, \mathbf{e}_1 \rangle^2}{\|\tilde{\boldsymbol{\phi}}_n\|^2}.$$

根据 Cauchy–Schwarz 不等式及 (5-8)，得

$$\begin{aligned}
\mathbb{E}[v(\boldsymbol{\phi}_n)|\boldsymbol{\phi}_0] \cdot \mathbb{E}\left[\|\tilde{\boldsymbol{\phi}}_n\|^2 \middle| \boldsymbol{\phi}_0\right] &\geqslant \left(\mathbb{E}\left[|\langle \tilde{\boldsymbol{\phi}}_n, \mathbf{e}_1 \rangle| \middle| \boldsymbol{\phi}_0\right]\right)^2 \\
&\geqslant \left(\mathbf{e}_1^\mathsf{T} \mathbb{E}\left[\tilde{\boldsymbol{\phi}}_n \middle| \boldsymbol{\phi}_0\right]\right)^2 \\
&= \left[\mathbf{e}_1^\mathsf{T} (1 + \eta \boldsymbol{\Sigma}^2)^n \tilde{\boldsymbol{\phi}}_0\right]^2 = (1 + \eta\sigma_1^2)^{2n} \mathbf{e}_1^\mathsf{T} \tilde{\boldsymbol{\phi}}_0^{\circ 2},
\end{aligned}$$

从而可知

$$\mathbb{E}[v(\boldsymbol{\phi}_n)|\boldsymbol{\phi}_0] \geqslant \frac{(1+\eta\sigma_1^2)^{2n}\mathbf{e}_1^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbb{E}[\|\tilde{\boldsymbol{\phi}}_n\|^2|\boldsymbol{\phi}_0]} = \frac{(1+\eta\sigma_1^2)^{2n}\mathbf{e}_1^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{\mathbf{1}^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}. \tag{C-1}$$

不失一般性，假设 $\|\boldsymbol{\phi}_0\|_2 = 1$，亦即，$\|\tilde{\boldsymbol{\phi}}_0^{\circ 2}\|_1 = 1$。注意到

$$\frac{\mathbf{1}^{\mathsf{T}}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{(1+\eta\sigma_1^2)^{2n}} = \frac{\mathbf{1}^{\mathsf{T}}\mathbf{Q}\boldsymbol{\Lambda}^n\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{(1+\eta\sigma_1^2)^{2n}} = \sum_{i=1}^{m}\frac{\lambda_i^n}{(1+\eta\sigma_1^2)^n}\cdot(\mathbf{1}^{\mathsf{T}}\mathbf{Q}\mathbf{e}_i\mathbf{e}_i^{\mathsf{T}}\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}), \tag{C-2}$$

最后乘积表达式中的两项可分别处理。

对于第一项，由 (5-14) 知存在 $\eta_1 > 0$，使得对任意 $\eta < \eta_1$，有 $\lambda_1 < (1+\eta\sigma_1^2)^2 + 2\eta^2\sigma_1^4\tau_1$，故

$$\begin{aligned}\log\frac{\lambda_1}{(1+\eta\sigma_1^2)^2} &\leqslant \log\frac{(1+\eta\sigma_1^2)^2 + 2\eta^2\sigma_1^4\tau_1}{(1+\eta\sigma_1^2)^2}\\ &\leqslant \frac{2\eta^2\sigma_1^4\tau_1}{(1+\eta\sigma_1^2)^2} < 2\eta^2\sigma_1^4\tau_1,\end{aligned}$$

其中第二个等号依据 $\log x \leqslant x-1$，对所有 $x > 0$。从而可得

$$\frac{\lambda_1^n}{(1+\eta\sigma_1^2)^{2n}} = \exp\left(n\cdot\log\frac{\lambda_1}{(1+\eta\sigma_1^2)^2}\right) \tag{C-3a}$$

$$\leqslant \exp\left(2n\eta^2\sigma_1^4\tau_1\right) \tag{C-3b}$$

$$\leqslant 1 + C_1\eta\log\frac{1}{\eta}, \tag{C-3c}$$

其中 $C_1 > 0$ 为常数。

与之类似，当 $i > 1$ 时存在 $C_2 > 0$ 与 $\eta_2 > 0$，使得对任意 $\eta < \eta_2$，有

$$\begin{aligned}\log\frac{\lambda_i}{(1+\eta\sigma_1^2)^2} &< 2\eta(\sigma_i^2 - \sigma_1^2) + C_2\eta^2,\\ &< 2\eta(\sigma_2^2 - \sigma_1^2) + C_2\eta^2,\end{aligned}$$

以及

$$\frac{\lambda_i^n}{(1+\eta\sigma_1^2)^{2n}} = \exp\left(n\cdot\log\frac{\lambda_i}{(1+\eta\sigma_1^2)^2}\right) \tag{C-4a}$$

$$< \exp\left(2n\eta(\sigma_2^2 - \sigma_1^2) + C_2n\eta^2\right) \tag{C-4b}$$

$$= \eta^2\exp(C_2n\eta^2) < 2\eta^2. \tag{C-4c}$$

对于第二项，由引理 5.1 的证明可知 $\mathbf{G}$ 的奇异向量为 η 的连续可导函数，从而 $\mathbf{Q}$ 的 Taylor 展开存在，且可表示为 $\mathbf{Q}=\mathbf{Q}_0+\eta\mathbf{Q}_1+o(\eta)$。此外由于 λ_1 为单特征值 (即代数重数为 1)，易知 $\mathbf{Q}_0$ 满足 $\mathbf{e}_1^\mathsf{T}\mathbf{Q}_0^{-1}=\mathbf{e}_1,\mathbf{Q}_0\mathbf{e}_1^\mathsf{T}=\mathbf{e}_1^\mathsf{T}$。因此由 Taylor 展开式可知

$$\begin{aligned}&\mathbf{1}^\mathsf{T}\mathbf{Q}\mathbf{e}_i\mathbf{e}_i^\mathsf{T}\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}\\&\quad=\mathbf{1}^\mathsf{T}\mathbf{Q}_0\mathbf{e}_i\mathbf{e}_i^\mathsf{T}\mathbf{Q}_0^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}-\eta\mathbf{1}^\mathsf{T}\mathbf{Q}_0\mathbf{e}_i\mathbf{e}_i^\mathsf{T}\mathbf{Q}_0^{-1}\mathbf{Q}_1\mathbf{Q}_0^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}+\eta\mathbf{1}^\mathsf{T}\mathbf{Q}_1\mathbf{e}_i\mathbf{e}_i^\mathsf{T}\mathbf{Q}_0^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}+o(\eta).\end{aligned}$$

故存在 $\eta_3>0,C_3>0$，使得对任意的 $\eta<\eta_3$，有

$$\mathbf{1}^\mathsf{T}\mathbf{Q}\mathbf{e}_1\mathbf{e}_1^\mathsf{T}\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}<\nu(\boldsymbol{\phi}_0)+\eta C_3,\tag{C-5a}$$

$$\mathbf{1}^\mathsf{T}\mathbf{Q}\mathbf{e}_i\mathbf{e}_i^\mathsf{T}\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2}<C_3,\quad i>1.\tag{C-5b}$$

将 (C-3)、(C-4)、以及 (C-5) 代入 (C-2)，可得

$$\begin{aligned}\frac{\mathbf{1}^\mathsf{T}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{(1+\eta\sigma_1^2)^{2n}}&=\sum_{i=1}^m\frac{\lambda_i^n}{(1+\eta\sigma_1^2)^n}\cdot(\mathbf{1}^\mathsf{T}\mathbf{Q}\mathbf{e}_i\mathbf{e}_i^\mathsf{T}\mathbf{Q}^{-1}\tilde{\boldsymbol{\phi}}_0^{\circ 2})\\&\leqslant\left(1+C_1\eta\log\frac{1}{\eta}\right)\cdot(\nu(\boldsymbol{\phi}_0)+\eta(C_3))+2(m-1)C_3\eta^2\end{aligned}$$

对任意 $\eta<\min\{\eta_1,\eta_2,\eta_3\}$ 成立。由于 $\eta=o(\eta\log\frac{1}{\eta}),\eta^2=o(\eta\log\frac{1}{\eta})$，存在 $\eta_4>0$ 及 $c>0$，使得对所有的 $\eta<\eta_4$，有

$$\frac{\mathbf{1}^\mathsf{T}\mathbf{G}^n\tilde{\boldsymbol{\phi}}_0^{\circ 2}}{(1+\eta\sigma_1^2)^{2n}}\leqslant\nu(\boldsymbol{\phi}_0)+c\eta\log\frac{1}{\eta}.$$

于是由 (C-1)，对 $\eta<\eta_0\triangleq\min\{\eta_1,\eta_2,\eta_3,\eta_4\}$，有

$$\begin{aligned}\mathbb{E}[\nu(\boldsymbol{\phi}_n)|\boldsymbol{\phi}_0]&\geqslant\nu(\boldsymbol{\phi}_0)\left(\nu(\boldsymbol{\phi}_0)+c\eta\log\frac{1}{\eta}\right)^{-1}\\&=1-\frac{c}{\nu(\boldsymbol{\phi}_0)}\cdot\eta\log\frac{1}{\eta}.\end{aligned}$$

根据 $\nu(\boldsymbol{\phi}_n)$ 的定义，$1-\nu(\boldsymbol{\phi}_n)$ 恒为非负。从而由 Markov 不等式可知

$$1-\nu(\boldsymbol{\phi}_n)\leqslant\frac{c}{\nu(\boldsymbol{\phi}_0)}\cdot\frac{\eta}{\delta}\log\frac{1}{\eta}$$

以不小于 $1-\delta$ 的概率成立。最后，(5-16) 可由 $\rho(\boldsymbol{\phi}_n)\geqslant\sigma_1^2\nu(\boldsymbol{\phi}_n)$ 推出。

C.3　命题 5.3 的证明

在该证明中，引入记号 $\mathbf{M}^{\otimes 2} \triangleq \mathbf{M} \otimes \mathbf{M}, \mathbf{M}^{\otimes 2\mathsf{T}} \triangleq (\mathbf{M}^{\otimes 2})^{\mathsf{T}}$，其中“ $\otimes$ ”表示矩阵间的 Kronecker 积。此外，对任意的 $i \leqslant j$，定义

$$\hat{\mathbf{e}}_{ij} \triangleq \begin{cases} \mathbf{e}_i^{\otimes 2}, & \text{if } i = j \\ \frac{1}{\sqrt{2}}(\mathbf{e}_i \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{e}_i), & \text{if } i < j \end{cases}$$

其中 $\mathbf{e}_i$ 表示 $\mathbb{R}^d$ 中的第 i 个标准基 (自然基) 向量。在此基础上，定义 $\mathbf{E} \in \mathbb{R}^{d^2 \times \tilde{d}}$ 为由所有 $\hat{\mathbf{e}}_{ij}$ $(i \leqslant j)$ 构成的矩阵，使其第 $[(i-1)i/2 + j]$ 列为 $\hat{\mathbf{e}}_{ij}$。定义 $\boldsymbol{\theta}_n \in \mathbb{R}^{\tilde{d}}$ 使其第 $[i + (j-1)j/2]$ $(i \leqslant j \leqslant d)$ 个元素为 $\boldsymbol{\phi}_n^{\mathsf{T}} \mathbf{V}_{ij} \boldsymbol{\phi}_n$，则易知 $\boldsymbol{\theta}_n = \mathbf{E}^{\mathsf{T}} \tilde{\boldsymbol{\phi}}_n^{\otimes 2}$。关于 $\mathbf{E}$，我们有如下引理。

引理 C.1： 对任意的 $\mathbf{u} \in \mathbb{R}^d$，有

$$\mathbf{E}\mathbf{E}^{\mathsf{T}} \mathbf{u}^{\otimes 2} = \mathbf{u}^{\otimes 2}. \tag{C-6}$$

证明 设 $\mathbf{u} = [u_1, \cdots, u_d]^{\mathsf{T}}$，则

$$\mathbf{u} = \sum_{i=1}^{d} u_i \mathbf{e}_i,$$

故有

$$\begin{aligned} \mathbf{u}^{\otimes 2} &= \left(\sum_{i=1}^{d} u_i \mathbf{e}_i \right) \otimes \left(\sum_{j=1}^{d} u_j \mathbf{e}_j \right) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} u_i u_j (\mathbf{e}_i \otimes \mathbf{e}_j) \\ &= \sum_{i=1}^{d} u_i^2 \mathbf{e}_i^{\otimes 2} + \sum_{i<j} u_i u_j (\mathbf{e}_i \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{e}_i) \\ &= \sum_{i=1}^{d} u_i^2 \hat{\mathbf{e}}_{ii} + \sqrt{2} \sum_{i<j} u_i u_j \hat{\mathbf{e}}_{ij}. \end{aligned} \tag{C-7}$$

此外，注意到由于

$$\mathbf{E}\mathbf{E}^{\mathsf{T}} = \sum_{i \leqslant j} \hat{\mathbf{e}}_{ij} \hat{\mathbf{e}}_{ij}^{\mathsf{T}},$$

可得

$$\begin{aligned}\mathbf{E}\mathbf{E}^{\mathsf{T}}\mathbf{u}^{\otimes 2} &= \sum_{i\leqslant j}\hat{\mathbf{e}}_{ij}\hat{\mathbf{e}}_{ij}^{\mathsf{T}}\mathbf{u}^{\otimes 2}\\ &= \sum_{i\leqslant j}\langle\hat{\mathbf{e}}_{ij},\mathbf{u}^{\otimes 2}\rangle\hat{\mathbf{e}}_{ij}\\ &= \sum_{i=1}^{d}u_i^2\hat{\mathbf{e}}_{ii}+\sqrt{2}\sum_{i<j}u_iu_j\hat{\mathbf{e}}_{ij}.\end{aligned}\tag{C-8}$$

综合 (C-7) 与 (C-8) 的结论可推出 (C-6)。 □

命题的证明还将用到如下关于 $\mathbf{L}$ 的引理。

引理 C.2： 由 (5-26) 所定义的矩阵 $\mathbf{L}$ 可表示为

$$\mathbf{L}=\mathbf{E}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\mathbf{E}.\tag{C-9}$$

证明 对任意的 $i\leqslant j$ 及 $i'\leqslant j'$，矩阵 $\mathbf{E}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\mathbf{E}$ 第 $[i+(j-1)j/2]$ 行第 $[i'+(j'-1)j'/2]$ 列的元素为

$$\begin{aligned}\mathbf{e}_{ij}^{\mathsf{T}}\mathbf{E}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\mathbf{E}\mathbf{e}_{i'j'} &= \hat{\mathbf{e}}_{ij}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\hat{\mathbf{e}}_{i'j'} &\text{(C-10)}\\ &= \mathrm{vec}^{\mathsf{T}}(\mathbf{V}_{ij})\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathrm{vec}(\mathbf{V}_{i'j'}) &\text{(C-11)}\\ &= \mathrm{vec}^{\mathsf{T}}(\mathbf{V}_{ij})\mathbb{E}\left[\mathrm{vec}(\mathbf{Z}\mathbf{V}_{i'j'}\mathbf{Z}^{\mathsf{T}})\right] &\text{(C-12)}\\ &= \mathbb{E}\left[\mathrm{tr}\left\{\mathbf{V}_{ij}\mathbf{Z}\mathbf{V}_{i'j'}\mathbf{Z}^{\mathsf{T}}\right\}\right] &\text{(C-13)}\\ &= L_{ij,i'j'} &\text{(C-14)}\end{aligned}$$

其中 $\mathrm{vec}(\cdot)$ 表示矩阵的向量化操作。在上述推导中，为导出 (C-11)，注意到对 $i\leqslant j$ 有 $\mathrm{vec}(\mathbf{V}_{ij})=\mathbf{V}^{\otimes 2}\hat{\mathbf{e}}_{ij}$；为导出 (C-13)，注意到对维度相容的矩阵 $\mathbf{M}_1$、$\mathbf{M}_2$，有 $\mathrm{vec}^{\mathsf{T}}(\mathbf{M}_1)\,\mathrm{vec}(\mathbf{M}_2)=\mathrm{tr}\left\{\mathbf{M}_1^{\mathsf{T}}\mathbf{M}_2\right\}$。故 $\mathbf{E}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\mathbf{E}$ 等于 $\mathbf{L}$。 □

基于前述引理，可得命题 5.3 的证明如下。

证明 (命题 5.3 的证明) 首先，令 $\tilde{\boldsymbol{\phi}}_n\triangleq\mathbf{V}^{\mathsf{T}}\boldsymbol{\phi}_n,\tilde{\boldsymbol{\xi}}_n\triangleq\mathbf{V}^{\mathsf{T}}\boldsymbol{\zeta}_n$, 则由 (5-21) 可知

$$\tilde{\boldsymbol{\phi}}_n=(\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1}+\eta\tilde{\boldsymbol{\xi}}_n,\quad i\geqslant 1,\tag{C-15}$$

由此推出

$$\tilde{\boldsymbol{\phi}}_n^{\otimes 2}=(\mathbf{I}+\eta\boldsymbol{\Sigma})^{\otimes 2}\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}+\eta((\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1})\otimes\tilde{\boldsymbol{\xi}}_n+\eta\tilde{\boldsymbol{\xi}}_n\otimes((\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1})+\eta^2\tilde{\boldsymbol{\xi}}_n^{\otimes 2}.\tag{C-16}$$

接下来考察 (C-16) 右边条件在 $\boldsymbol{\phi}_0$ 上的条件期望。由于噪声 $\mathbf{Z}_n$ 互相独立，知 $\boldsymbol{\phi}_0 \leftrightarrow \boldsymbol{\phi}_1 \leftrightarrow \cdots \leftrightarrow \boldsymbol{\phi}_n$ 构成 Markov 链，从而

$$\begin{aligned}\mathbb{E}\big[((\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1})\otimes\tilde{\boldsymbol{\xi}}_n\big|\boldsymbol{\phi}_0\big] &= \mathbb{E}\big[\mathbb{E}\big[((\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1})\otimes\tilde{\boldsymbol{\xi}}_n\big|\boldsymbol{\phi}_0,\boldsymbol{\phi}_{n-1}\big]\big|\boldsymbol{\phi}_0\big]\\ &= \mathbb{E}\big[((\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1})\otimes\mathbb{E}\big[\tilde{\boldsymbol{\xi}}_n\big|\boldsymbol{\phi}_{n-1}\big]\big|\boldsymbol{\phi}_0\big]\\ &= \mathbf{0}^{\otimes 2},\end{aligned}$$

其中最后的等号基于 $\mathbb{E}\big[\tilde{\boldsymbol{\xi}}_n\big|\boldsymbol{\phi}_{n-1}\big] = \mathbf{V}^{\mathsf{T}}\mathbb{E}\big[\boldsymbol{\zeta}_n\big|\boldsymbol{\phi}_{n-1}\big] = \mathbb{E}\left[\mathbf{Z}_n\right]\boldsymbol{\phi}_{n-1} = \mathbf{0}$。同法可得 $\mathbb{E}\big[\tilde{\boldsymbol{\xi}}_n\otimes((\mathbf{I}+\eta\boldsymbol{\Sigma})\tilde{\boldsymbol{\phi}}_{n-1})\big|\boldsymbol{\phi}_0\big] = \mathbf{0}^{\otimes 2}$。

此外，注意到由

$$\tilde{\boldsymbol{\xi}}_n = \mathbf{V}^{\mathsf{T}}\boldsymbol{\zeta}_n = \mathbf{V}^{\mathsf{T}}\mathbf{Z}_n\boldsymbol{\phi}_{n-1} = \mathbf{V}^{\mathsf{T}}\mathbf{Z}_n\mathbf{V}\tilde{\boldsymbol{\phi}}_{n-1}$$

可推出

$$\begin{aligned}\mathbb{E}\left[\tilde{\boldsymbol{\xi}}_n^{\otimes 2}\middle|\boldsymbol{\phi}_0\right] &= \mathbb{E}\left[\mathbb{E}\left[\tilde{\boldsymbol{\xi}}_n^{\otimes 2}\middle|\boldsymbol{\phi}_0,\boldsymbol{\phi}_{n-1}\right]|\boldsymbol{\phi}_0\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\tilde{\boldsymbol{\xi}}_n^{\otimes 2}\middle|\boldsymbol{\phi}_{n-1}\right]|\boldsymbol{\phi}_0\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[(\mathbf{V}^{\mathsf{T}}\mathbf{Z}_n\mathbf{V}\tilde{\boldsymbol{\phi}}_{n-1})^{\otimes 2}\middle|\boldsymbol{\phi}_{n-1}\right]\middle|\boldsymbol{\phi}_0\right]\\ &= \mathbb{E}\left[\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}_n^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\middle|\boldsymbol{\phi}_0\right]\\ &= \mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\mathbb{E}\left[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\middle|\boldsymbol{\phi}_0\right]\end{aligned}\tag{C-17}$$

其中第二个等号利用了 $\boldsymbol{\phi}_0 - \boldsymbol{\phi}_1 - \cdots - \boldsymbol{\phi}_n$ 构成 Markov 链的性质。

对 (C-16) 取关于 $\boldsymbol{\phi}_0$ 的条件期望，得

$$\mathbb{E}\left[\tilde{\boldsymbol{\phi}}_n^{\otimes 2}\middle|\boldsymbol{\phi}_0\right] = \left[(\mathbf{I}+\eta\boldsymbol{\Sigma})^{\otimes 2} + \eta^2\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\right]\mathbb{E}\left[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\middle|\boldsymbol{\phi}_0\right],$$

因此

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{\theta}_n\big|\boldsymbol{\phi}_0\big] &= \mathbb{E}\big[\mathbf{E}^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_n^{\otimes 2}\big|\boldsymbol{\phi}_0\big]\\ &= \mathbf{E}^{\mathsf{T}}\left[(\mathbf{I}+\eta\boldsymbol{\Sigma})^{\otimes 2} + \eta^2\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\right]\mathbb{E}\big[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\big|\boldsymbol{\phi}_0\big]\\ &= \left(\mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Sigma}_1 + \eta^2\boldsymbol{\Sigma}_2\right)\mathbf{E}^{\mathsf{T}}\mathbb{E}\big[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\big|\boldsymbol{\phi}_0\big]\\ &\quad + \eta^2\mathbf{E}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\cdot\mathbb{E}\big[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\big|\boldsymbol{\phi}_0\big] \qquad \text{(C-18)}\\ &= \left(\mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Sigma}_1 + \eta^2\boldsymbol{\Sigma}_2\right)\mathbf{E}^{\mathsf{T}}\mathbb{E}\big[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\big|\boldsymbol{\phi}_0\big]\\ &\quad + \eta^2\mathbf{E}^{\mathsf{T}}\mathbf{V}^{\otimes 2\mathsf{T}}\mathbb{E}\left[\mathbf{Z}^{\otimes 2}\right]\mathbf{V}^{\otimes 2}\mathbf{E}\mathbf{E}^{\mathsf{T}}\mathbb{E}\big[\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\big|\boldsymbol{\phi}_0\big] \qquad \text{(C-19)}\\ &= (\mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Sigma}_1 + \eta^2\boldsymbol{\Sigma}_2 + \eta^2\mathbf{L})\mathbb{E}\big[\mathbf{E}^{\mathsf{T}}\tilde{\boldsymbol{\phi}}_{n-1}^{\otimes 2}\big|\boldsymbol{\phi}_0\big] \qquad \text{(C-20)}\end{aligned}$$

$$= \mathbf{G}\mathbb{E}\big[\boldsymbol{\theta}_{n-1}\big|\boldsymbol{\phi}_0\big], \tag{C-21}$$

其中 (C-18) 基于 $\mathbf{E}^\top(\mathbf{I}+\eta\boldsymbol{\Sigma})^{\otimes 2} = (\mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Sigma}_1 + \eta^2\boldsymbol{\Sigma}_2)\mathbf{E}^\top$，(C-19) 基于 引理 C.1, (C-20) 基于 引理 C.2，且 (C-21) 基于 $\mathbf{G}$ 的定义。

由此得出

$$\mathbb{E}\big[\boldsymbol{\theta}_n\big|\boldsymbol{\phi}_0\big] = \mathbf{G}\mathbb{E}\big[\boldsymbol{\theta}_{n-1}\big|\boldsymbol{\phi}_0\big] = \cdots = \mathbf{G}^n\boldsymbol{\theta}_0$$

从而有

$$\begin{aligned}\pi_n^{(i)}(\boldsymbol{\phi}_0) = \mathbb{E}\big[\langle\boldsymbol{\phi}_n, \mathbf{v}_i\rangle^2\big|\boldsymbol{\phi}_0\big] &= \mathbb{E}\big[\langle\mathbf{e}_{ii}, \boldsymbol{\theta}_n\rangle\big|\boldsymbol{\phi}_0\big] \\ &= \langle\mathbf{e}_{ii}, \mathbb{E}\big[\boldsymbol{\theta}_n\big|\boldsymbol{\phi}_0\big]\rangle \\ &= \langle\mathbf{e}_{ii}, \mathbf{G}^n\boldsymbol{\theta}_0\rangle.\end{aligned}$$

□

C.4　引理 5.2 的证明

证明 令 $\mathbf{G}_0 \triangleq \boldsymbol{\Sigma}_1 + \eta(\boldsymbol{\Sigma}_2 + \mathbf{L})$，则 $\mathbf{G} = \mathbf{I}_{\tilde{d}} + \eta\mathbf{G}_0$，故 $\mathbf{G}$ 的特征值分解可立即由 $\mathbf{G}_0$ 的特征值分解得到。基于有关 σ_i 的假设可知 $\boldsymbol{\Sigma}_1$ 各对角元相异，因此对充分小的 η，矩阵 $\mathbf{G}_0 = \boldsymbol{\Sigma}_1 + \eta(\boldsymbol{\Sigma}_2 + \mathbf{L})$ 可对角化。记 $\mathbf{G}_0$ 的特征值分解为

$$\mathbf{G}_0 = \mathbf{Q}(\eta)\boldsymbol{\Lambda}(\eta)\mathbf{Q}(\eta)^{-1}, \tag{C-22}$$

并记 $\lambda_{ij}(\mathbf{G}_0)$ $(i \leqslant j)$ 为相应的特征值，即 $\boldsymbol{\Lambda}$ 的对角元。注意到 $\mathbf{G}_0$ 可视为 $\boldsymbol{\Sigma}_1$ 扰动之后的矩阵，根据微扰论[96] 可知扰动之后的特征值及特征向量分别为

$$\lambda_{ij}(\mathbf{G}_0) = \sigma_i + \sigma_j + \eta(\sigma_i\sigma_j + L_{ij,ij}) + o(\eta), \quad \forall\, i \leqslant j$$

及 $\mathbf{Q} = \mathbf{I}_{\tilde{d}} + \eta\mathbf{Q}_1 + o(\eta)$，其中 $\mathbf{Q}_1$ 满足

$$\mathbf{e}_{ii}^\top\mathbf{Q}_1\mathbf{e}_{i'j'} = \frac{L_{ii,i'j'}}{\sigma_i' + \sigma_j' - 2\sigma_i}, \quad \forall\,(i', j') \neq (i, i). \tag{C-23}$$

由 (C-22) 可知

$$\mathbf{G}^n = [\mathbf{Q}(\eta)(\mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Lambda}(\eta))\mathbf{Q}(\eta)^{-1}]^n = \mathbf{Q}(\eta)(\mathbf{I}_{\tilde{d}} + \eta\boldsymbol{\Lambda}(\eta))^n\mathbf{Q}(\eta)^{-1},$$

且 $\mathbf{G}$ 特征值为

$$\lambda_{ij}(\mathbf{G}) = 1 + \eta\lambda_{ij}(\mathbf{G}_0) = 1 + \eta(\sigma_i + \sigma_j) + \eta^2(\sigma_i\sigma_j + L_{ij,ij}) + o(\eta^2).$$

从而可得

$$\begin{aligned}\frac{1}{(\lambda_{11}(\mathbf{G}))^n}\mathbf{G}^n &= \frac{1}{(\lambda_{11}(\mathbf{G}))^n}\mathbf{Q}(\eta)(\mathbf{I}_{\tilde{d}}+\eta\mathbf{\Lambda}(\eta))^n\mathbf{Q}(\eta)^{-1}\\ &= (\mathbf{I}_{\tilde{d}}+\eta\mathbf{Q}_1)\left(\frac{\mathbf{I}_{\tilde{d}}+\eta\mathbf{\Lambda}(\eta)}{\lambda_{11}(\mathbf{G})}\right)^n(\mathbf{I}_{\tilde{d}}-\eta\mathbf{Q}_1)+o(\eta)\\ &= \sum_{i'\leqslant j'}\left[\frac{\lambda_{i'j'}(\mathbf{G})}{\lambda_{11}(\mathbf{G})}\right]^n(\mathbf{I}_{\tilde{d}}+\eta\mathbf{Q}_1)\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top(\mathbf{I}_{\tilde{d}}-\eta\mathbf{Q}_1)+o(\eta)\\ &= \sum_{i'\leqslant j'}\gamma_{i'j'}^n(\eta)\left[\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top+\eta(\mathbf{Q}_1\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top-\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top\mathbf{Q}_1)\right]+o(\eta),\end{aligned}$$

其中第二个等式依据 $\mathbf{Q}^{-1}=\mathbf{I}_{\tilde{d}}-\eta\mathbf{Q}_1+o(\eta)$。

因此可得

$$\begin{aligned}\frac{\mathbf{e}_{ii}^\top\mathbf{G}^n}{(\lambda_{11}(\mathbf{G}))^n} &= \sum_{i'\leqslant j'}\gamma_{i'j'}^n(\eta)\cdot\mathbf{e}_{ii}^\top\left[\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top+\eta(\mathbf{Q}_1\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top-\mathbf{e}_{i'j'}\mathbf{e}_{i'j'}^\top\mathbf{Q}_1)\right]+o(\eta)\\ &= \gamma_{ii}^n(\eta)\left(\mathbf{e}_{ii}^\top-\eta\mathbf{e}_{ii}^\top\mathbf{Q}_1\right)+\eta\sum_{i'\leqslant j'}\gamma_{i'j'}^n(\eta)(\mathbf{e}_{ii}^\top\mathbf{Q}_1\mathbf{e}_{i'j'})\mathbf{e}_{i'j'}^\top+o(\eta)\\ &= \gamma_{ii}^n(\eta)\mathbf{e}_{ii}^\top+\eta\sum_{i'\leqslant j'}[\gamma_{i'j'}^n(\eta)-\gamma_{ii}^n(\eta)](\mathbf{e}_{ii}^\top\mathbf{Q}_1\mathbf{e}_{i'j'})\mathbf{e}_{i'j'}^\top+o(\eta)\\ &= \gamma_{ii}^n(\eta)\mathbf{e}_{ii}^\top+\eta\sum_{\substack{i'\leqslant j'\\(i',j')\neq(i,i)}}\frac{\gamma_{i'j'}^n(\eta)-\gamma_{ii}^n(\eta)}{\sigma_{i'}+\sigma_{j'}-2\sigma_i}L_{ii,i'j'}\cdot\mathbf{e}_{i'j'}^\top+o(\eta),\end{aligned}$$

其中最后的等号依据 (C-23)。最后，将上式两边与 $\boldsymbol{\theta}_0$ 作内积即得 (5-33)。 □

C.5　定理 5.5 的证明

为刻画大样本小学习率机制下泛化误差的行为，引入记号 $f(\eta)=o^{(\eta)}(1)$ 以表示 $\lim_{\eta\to0^+}f(\eta)=0$，并用 $g(\eta,N)=o^{(N\eta)}(1)$ 表示存在函数 $h(t)$ 使得

$$\lim_{t\to+\infty}h(t)=0 \text{ 以及 } |g(\eta,N)|\leqslant|h(N\eta)|,$$

则易验证 $\hat{o}(1)=o^{(N\eta)}(1)+o^{(\eta)}(1)$。

此外，如下引理将用于后续证明中。

引理 C.3： 我们有

$$\gamma_{ij}^N(\eta)=o^{(N\eta)}(1),\quad (i,j)\neq(1,1), \tag{C-24a}$$

$$\gamma_{22}^N(\eta)=e^{2(\sigma_2-\sigma_1)N\eta}(1+o^{(N\eta)}(1))+\eta\cdot o^{(\eta)}(1), \tag{C-24b}$$

$$\gamma_{ii}^N(\eta) = \gamma_{22}^N(\eta) \cdot o^{(N\eta)}(1), \quad i > 2. \tag{C-24c}$$

证明 根据 (5.2) 可得

$$\gamma_{ij}(\eta) = \frac{\lambda_{ij}(\mathbf{G})}{\lambda_{11}(\mathbf{G})} = 1 + \eta(\sigma_i + \sigma_j - 2\sigma_1) + o(\eta), \tag{C-25}$$

故对 $i > 1$ 及充分小的 η 有

$$\gamma_{ii}(\eta) = 1 + 2\eta(\sigma_i - \sigma_1) + o(\eta) \tag{C-26a}$$

$$< 1 + \eta(\sigma_i - \sigma_1) \tag{C-26b}$$

$$\leqslant \exp(\eta(\sigma_i - \sigma_1)) \tag{C-26c}$$

从而

$$\gamma_{ii}^N(\eta) < \exp(N\eta(\sigma_i - \sigma_1)),$$

其中 (C-26c) 基于不等式 $1 + x \leqslant e^x$。

对任意 $(i, j) \neq (1, 1)$，由 (C-25) 可知

$$\gamma_{ij}(\eta) = 1 + \eta(\sigma_i + \sigma_j - 2\sigma_1) + o(\eta) \tag{C-27}$$

$$< 1 + \eta\left(\frac{\sigma_i + \sigma_j}{2} - \sigma_1\right) \tag{C-28}$$

$$\leqslant \exp\left(\eta\left(\frac{\sigma_i + \sigma_j}{2} - \sigma_1\right)\right) \tag{C-29}$$

$$\leqslant \exp\left(\eta \cdot \frac{\sigma_2 - \sigma_1}{2}\right) \tag{C-30}$$

其中 (C-29) 使用了不等式 $1 + x \leqslant e^x$。因此可得

$$\gamma_{ij}^N(\eta) < \exp\left(N\eta \cdot \frac{\sigma_2 - \sigma_1}{2}\right),$$

由此推出 (C-24a)。

为证明 (C-24b)，首先注意到根据 (5-32) 有

$$\begin{aligned}\log \lambda_{ij}(\mathbf{G}) &= \eta(\sigma_i + \sigma_j) + \eta^2(\sigma_i\sigma_j + L_{ij,ij}) - \frac{1}{2}\eta^2(\sigma_i + \sigma_j)^2 + o(\eta^2) \\ &= \eta(\sigma_i + \sigma_j) + \eta^2\left[L_{ij,ij} - \frac{1}{2}(\sigma_i^2 + \sigma_j^2)\right] + o(\eta^2).\end{aligned}$$

因此，根据定义 (5-34) 有

$$\begin{aligned}\gamma_{ij}(\eta) = \frac{\lambda_{ij}(\mathbf{G})}{\lambda_{11}(\mathbf{G})} &= \exp\left(\log\lambda_{ij}(\mathbf{G}) - \log\lambda_{11}(\mathbf{G})\right) \\ &= \exp\left(\eta(\sigma_i+\sigma_j-2\sigma_1) + c_{ij}\eta^2 + o(\eta^2)\right),\end{aligned}$$

其中 $c_{ij} = \sigma_1^2 - \frac{\sigma_i^2+\sigma_j^2}{2} + (L_{ij,ij} - L_{11,11})$。特别地，可得 $\gamma_{22}(\eta) = \exp\left(2\eta(\sigma_2-\sigma_1) + c_{22}\eta^2 + o(\eta^2)\right)$，故存在常数 $C>0$ 使得对充分小的 η，有

$$\exp\left(2\eta(\sigma_2-\sigma_1) - C\eta^2\right) < \gamma_{22}(\eta) < \exp\left(2\eta(\sigma_2-\sigma_1) + C\eta^2\right).$$

于是可得出

$$(e^{-CN\eta^2}-1)\cdot e^{2N\eta(\sigma_2-\sigma_1)} < \gamma_{22}^N(\eta) - e^{2N\eta(\sigma_2-\sigma_1)} < (e^{CN\eta^2}-1)\cdot e^{2N\eta(\sigma_2-\sigma_1)},$$

因此

$$\left|\gamma_{22}^N(\eta) - e^{2N\eta(\sigma_2-\sigma_1)}\right| < (e^{CN\eta^2}-1)\cdot e^{2N\eta(\sigma_2-\sigma_1)}. \tag{C-31}$$

注意到

$$\begin{aligned}(e^{CN\eta^2}-1)\cdot e^{2N\eta(\sigma_2-\sigma_1)} &< CN\eta^2\cdot e^{CN\eta^2}\cdot e^{2N\eta(\sigma_2-\sigma_1)} && \text{(C-32)}\\ &= CN\eta^2\cdot e^{\frac{2}{3}N\eta(\sigma_2-\sigma_1)+CN\eta^2}\cdot e^{\frac{4}{3}N\eta(\sigma_2-\sigma_1)} \\ &< CN\eta^2\cdot e^{\frac{1}{3}N\eta(\sigma_2-\sigma_1)}\cdot e^{\frac{4}{3}N\eta(\sigma_2-\sigma_1)} && \text{(C-33)}\\ &= \eta e^{\frac{4}{3}N\eta(\sigma_2-\sigma_1)}\cdot\left[CN\eta\cdot e^{\frac{1}{3}N\eta(\sigma_2-\sigma_1)}\right] \\ &< \eta e^{\frac{4}{3}N\eta(\sigma_2-\sigma_1)}\cdot o^{(N\eta)}(1) && \text{(C-34)}\\ &\leqslant \frac{1}{2}\left(\eta^2 + e^{\frac{8}{3}N\eta(\sigma_2-\sigma_1)}\cdot o^{(N\eta)}(1)\right) && \text{(C-35)}\\ &= \eta\cdot o^{(\eta)}(1) + e^{2(\sigma_2-\sigma_1)N\eta}\cdot o^{(N\eta)}(1), && \text{(C-36)}\end{aligned}$$

其中 (C-32) 依据 $e^x - 1 \leqslant xe^x$，(C-33) 基于 η 很小的假设，(C-35) 基于算术–几何平均不等式。

从而 (C-24b) 可由 (C-31) 及 (C-36) 直接推出。

最后，为证明 (C-24c)，注意到由 (C-26a) 可知，对任意 $i>2$ 有

$$\frac{\gamma_{ii}(\eta)}{\gamma_{22}(\eta)} = 1 + 2\eta(\sigma_i-\sigma_2) + o(\eta) < 1 + \eta(\sigma_i-\sigma_2) < \exp(\eta(\sigma_i-\sigma_2)),$$

从而

$$0 < \frac{\gamma_{ii}^N(\eta)}{\gamma_{22}^N(\eta)} < \exp(-N\eta(\sigma_2 - \sigma_i)) \leqslant \exp(-N\eta(\sigma_2 - \sigma_3)),$$

因此 (C-24c) 成立。 □

在此基础上可证明定理 5.5 如下。

证明 (定理 5.5 的证明) 首先考察 $\bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)$。基于 (5-29) 的结果，只需考察 $\pi_N^{(i)}(\boldsymbol{\phi}_0)$。为此，注意到利用 引理 5.2 有

$$\frac{\pi_N^{(i)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N} = \eta \sum_{\substack{i' \leqslant j' \\ (i',j') \neq (i,i)}} \frac{\gamma_{i'j'}^N(\eta) - \gamma_{ii}^N(\eta)}{\sigma_{i'} + \sigma_{j'} - 2\sigma_i} \cdot L_{ii,i'j'} \langle \mathbf{e}_{i'j'}, \theta_0 \rangle + \gamma_{ii}^N(\eta) \langle \mathbf{e}_{ii}, \theta_0 \rangle + \eta o^{(\eta)}(1).$$

当 $i = 1$ 时，由 (C-24a) 可得

$$\frac{\pi_N^{(1)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N} = \langle \mathbf{e}_{11}, \theta_0 \rangle + \eta\alpha + \eta\hat{o}(1), \tag{C-37}$$

从而有

$$\frac{(\lambda_{11}(\mathbf{G}))^N}{\pi_N^{(1)}(\boldsymbol{\phi}_0)} = \frac{1}{\langle \mathbf{e}_{11}, \theta_0 \rangle} \left(1 - \frac{\eta}{\langle \mathbf{e}_{11}, \theta_0 \rangle} \alpha\right) + \eta\hat{o}(1), \tag{C-38}$$

其中 α 定义为

$$\alpha \triangleq \sum_{\substack{i' \leqslant j' \\ (i',j') \neq (1,1)}} \frac{L_{11,i'j'} \langle \mathbf{e}_{i'j'}, \theta_0 \rangle}{2\sigma_1 - \sigma_{i'} - \sigma_{j'}}.$$

类似地，当 $i > 1$ 时，有

$$\frac{\pi_N^{(i)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N} = \gamma_{ii}^N(\eta) \langle \mathbf{e}_{ii}, \theta_0 \rangle + \frac{\eta}{2} \cdot \frac{\tau_i}{\sigma_1 - \sigma_i} \langle \mathbf{e}_{11}, \theta_0 \rangle + \eta\hat{o}(1).$$

因此，

$$\begin{aligned} \sum_{i=2}^d \frac{\pi_N^{(i)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N} &= \gamma_{22}^N(\eta) \langle \mathbf{e}_{22}, \theta_0 \rangle (1 + o^{(N\eta)}(1)) + \frac{\eta}{2} \cdot \langle \mathbf{e}_{11}, \theta_0 \rangle \sum_{i=2}^d \frac{\tau_i}{\sigma_1 - \sigma_i} + \eta\hat{o}(1) \\ &= e^{2(\sigma_2 - \sigma_1)N\eta} \langle \mathbf{e}_{22}, \theta_0 \rangle (1 + o^{(N\eta)}(1)) + \frac{\eta}{2} \cdot \langle \mathbf{e}_{11}, \theta_0 \rangle \sum_{i=2}^d \frac{\tau_i}{\sigma_1 - \sigma_i} + \eta\hat{o}(1) \end{aligned} \tag{C-39}$$

其中第一个等号依据 (C-24c)，第二个等号基于 (C-24b)。

在此基础上，由 (C-38) 及 (C-39) 可推出

$$\begin{aligned}\frac{1}{\pi_n^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\pi_n^{(i)}(\boldsymbol{\phi}_0) &= \frac{(\lambda_{11}(\mathbf{G}))^N}{\pi_n^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\frac{\pi_n^{(i)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N}\\ &= e^{2(\sigma_2-\sigma_1)N\eta}\cdot\frac{\langle\mathbf{e}_{22},\boldsymbol{\theta}_0\rangle}{\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}(1+\hat{o}(1))+\frac{\eta}{2}\sum_{i=2}^{d}\frac{\tau_i}{\sigma_1-\sigma_i}(1+\hat{o}(1))\\ &= \left[e^{2(\sigma_2-\sigma_1)N\eta}\cdot\frac{\langle\mathbf{v}_2,\boldsymbol{\phi}_0\rangle^2}{\langle\mathbf{v}_1,\boldsymbol{\phi}_0\rangle^2}+\frac{\eta}{2}\sum_{i=2}^{d}\frac{\tau_i}{\sigma_1-\sigma_i}\right]\cdot(1+\hat{o}(1))\\ &= \hat{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)\cdot(1+\hat{o}(1))\end{aligned}\tag{C-40}$$

其中最后的等号成立基于 (C-24b)。

最后，由 (5-29) 可知

$$\bar{v}_N(\boldsymbol{\phi}_0)=\left(1+\frac{1}{\pi_n^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\pi_n^{(i)}(\boldsymbol{\phi}_0)\right)^{-1}=1-\hat{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)\cdot(1+\hat{o}(1))$$

故

$$\bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)=\hat{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)\cdot(1+\hat{o}(1)).\tag{C-41}$$

同法可得 $\bar{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)$ 的结果。具体地，由 (5-29) 可得

$$\begin{aligned}\bar{\rho}_N(\boldsymbol{\phi}_0) &= \left(\sigma_1+\frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\sigma_i\pi_N^{(i)}(\boldsymbol{\phi}_0)\right)\left(1+\frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\pi_N^{(i)}(\boldsymbol{\phi}_0)\right)^{-1}\\ &= \left(\sigma_1+\frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\sigma_i\pi_N^{(i)}(\boldsymbol{\phi}_0)\right)\bar{v}_N(\boldsymbol{\phi}_0).\end{aligned}\tag{C-42}$$

通过类似的推导可得 [对比 (C-40)]

$$\frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\sigma_i\pi_N^{(i)}(\boldsymbol{\phi}_0)=\left[\sigma_2 e^{2(\sigma_2-\sigma_1)N\eta}\cdot\frac{\langle\mathbf{v}_2,\boldsymbol{\phi}_0\rangle^2}{\langle\mathbf{v}_1,\boldsymbol{\phi}_0\rangle^2}+\frac{\eta}{2}\sum_{i=2}^{d}\frac{\sigma_i\tau_i}{\sigma_1-\sigma_i}\right]\cdot(1+\hat{o}(1)),\tag{C-43}$$

于是 (C-42) 可表示为

$$\bar{\rho}_N(\boldsymbol{\phi}_0)=\left(\sigma_1+\frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)}\sum_{i=2}^{d}\sigma_i\pi_N^{(i)}(\boldsymbol{\phi}_0)\right)\bar{v}_N(\boldsymbol{\phi}_0)$$

$$
\begin{aligned}
&= \left(\sigma_1 + \frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)} \sum_{i=2}^{d} \sigma_i \pi_N^{(i)}(\boldsymbol{\phi}_0)\right)\left[1 - \bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0)\right] \\
&= \sigma_1 - \sigma_1 \bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0) + \frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)} \sum_{i=2}^{d} \sigma_i \pi_N^{(i)}(\boldsymbol{\phi}_0) + (\eta + e^{2(\sigma_2-\sigma_1)N\eta})\hat{o}(1), \qquad \text{(C-44)} \\
&= \sigma_1 + \left[(\sigma_2 - \sigma_1)e^{2(\sigma_2-\sigma_1)N\eta} \cdot \frac{\langle \mathbf{v}_2, \boldsymbol{\phi}_0\rangle^2}{\langle \mathbf{v}_1, \boldsymbol{\phi}_0\rangle^2} - \frac{\eta}{2}\sum_{i=2}^{d} \tau_i\right] \cdot (1 + \hat{o}(1)) \\
&= \sigma_1 - \hat{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0) \cdot (1 + \hat{o}(1)) \qquad \text{(C-45)}
\end{aligned}
$$

其中 (C-44) 依据 [参见 (C-41) 及 (C-43)]

$$
\left(\frac{1}{\pi_N^{(1)}(\boldsymbol{\phi}_0)} \sum_{i=2}^{d} \sigma_i \pi_N^{(i)}(\boldsymbol{\phi}_0)\right) \bar{v}_N^{\mathsf{c}}(\boldsymbol{\phi}_0) = (e^{2(\sigma_2-\sigma_1)N\eta} + \eta)\hat{o}(1).
$$

因此得到

$$
\bar{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0) = \sigma_1 - \bar{\rho}_N(\boldsymbol{\phi}_0) = \hat{\rho}_N^{\mathsf{c}}(\boldsymbol{\phi}_0) \cdot (1 + \hat{o}(1)). \qquad \square
$$

C.6　定理 5.6 的证明

证明　由 (5-21) 可得 $\mathbb{E}\left[\boldsymbol{\phi}_n | \boldsymbol{\phi}_{n-1}\right] = (\mathbf{I} + \eta\mathbf{A})\boldsymbol{\phi}_{n-1}$，于是根据 $\boldsymbol{\phi}_n$ 满足的 Markov 性可知

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{\phi}_n | \boldsymbol{\phi}_0\right] &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\phi}_n | \boldsymbol{\phi}_0, \boldsymbol{\phi}_{n-1}\right] | \boldsymbol{\phi}_0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\phi}_n | \boldsymbol{\phi}_{n-1}\right] | \boldsymbol{\phi}_0\right] \\
&= (\mathbf{I} + \eta\mathbf{A})\mathbb{E}\left[\boldsymbol{\phi}_{n-1} | \boldsymbol{\phi}_0\right] \\
&= (\mathbf{I} + \eta\mathbf{A})^n \boldsymbol{\phi}_0.
\end{aligned}
$$

上式两边分别与 $\mathbf{v}_1$ 作内积，得

$$
\mathbb{E}\left[\langle \mathbf{v}_1, \boldsymbol{\phi}_n\rangle | \boldsymbol{\phi}_0\right] = (1 + \eta\sigma_1)^n \langle \mathbf{v}_1, \boldsymbol{\phi}_0\rangle.
$$

从而得到

$$
\begin{aligned}
\mathbb{E}\left[\nu(\boldsymbol{\phi}_N; \eta(N)) | \boldsymbol{\phi}_0\right] \cdot \mathbb{E}\left[\|\boldsymbol{\phi}_N\|^2 \middle| \boldsymbol{\phi}_0\right] &\geqslant \left(\mathbb{E}\left[|\langle \boldsymbol{\phi}_N, \mathbf{v}_1\rangle| \middle| \boldsymbol{\phi}_0\right]\right)^2 \\
&\geqslant \left(\mathbb{E}\left[\langle \boldsymbol{\phi}_N, \mathbf{v}_1\rangle \middle| \boldsymbol{\phi}_0\right]\right)^2 \\
&= (1 + \eta\sigma_1)^{2N} \langle \mathbf{v}_1, \boldsymbol{\phi}_0\rangle^2,
\end{aligned}
$$

其中两个不等号分别依据 Cauchy–Schwarz 不等式及 Jensen 不等式。由此得平均泛化误差 $\mathbb{E}\left[\nu(\boldsymbol{\phi}_N)|\boldsymbol{\phi}_0\right]$ 的下界

$$\mathbb{E}\left[\nu(\boldsymbol{\phi}_N)|\boldsymbol{\phi}_0\right] \geqslant \frac{(1+\eta\sigma_1)^{2N}\cdot\langle\mathbf{v}_1,\boldsymbol{\phi}_0\rangle^2}{\mathbb{E}\left[\|\boldsymbol{\phi}_N\|^2|\boldsymbol{\phi}_0\right]} \tag{C-46a}$$

$$= \bar{\nu}_N(\boldsymbol{\phi}_0)\cdot\frac{(1+\eta\sigma_1)^{2N}\cdot\langle\mathbf{v}_1,\boldsymbol{\phi}_0\rangle^2}{\mathbb{E}\left[\langle\mathbf{v}_1,\boldsymbol{\phi}_n\rangle^2|\boldsymbol{\phi}_0\right]} \tag{C-46b}$$

$$= \bar{\nu}_N(\boldsymbol{\phi}_0)\cdot\frac{(1+\eta\sigma_1)^{2N}\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}{\pi_N^{(1)}(\boldsymbol{\phi}_0)}. \tag{C-46c}$$

为进一步考察该下界，注意到

$$\begin{aligned}\frac{(1+\eta\sigma_1)^{2N}\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}{\pi_N^{(1)}(\boldsymbol{\phi}_0)} &= \left[\frac{\pi_N^{(1)}(\boldsymbol{\phi}_0)}{(1+\eta\sigma_1)^{2N}\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}\right]^{-1}\\ &= \left[\frac{(\lambda_{11}(\mathbf{G}))^N}{(1+\eta\sigma_1)^{2N}}\cdot\frac{\pi_N^{(1)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}\right]^{-1},\end{aligned}$$

乘积中的两项可分别处理。对第一项，注意到根据引理 5.2 有 $\lambda_{11}(\mathbf{G}) = (1+\eta\sigma_1)^2 + \eta^2\tau_1 + o(\eta^2)$，从而

$$\frac{\lambda_{11}(\mathbf{G})}{(1+\eta\sigma_1)^2} = 1 + \frac{\eta^2\tau_1+o(\eta^2)}{(1+\eta\sigma_1)^2} = 1+\eta^2\tau_1+o(\eta^2).$$

因此有

$$\begin{aligned}\left[\frac{\lambda_{11}(\mathbf{G})}{(1+\eta\sigma_1)^2}\right]^N &= \exp\left(N\log\frac{\lambda_{11}(\mathbf{G})}{(1+\eta\sigma_1)^2}\right)\\ &= \exp\left(N\log\left(1+\eta^2\tau_1+o(\eta^2)\right)\right)\\ &= \exp\left(N\eta^2\tau_1+o(N\eta^2)\right)\\ &= 1+N\eta^2\tau_1+o(N\eta^2).\end{aligned}$$

对于第二项，由 $\eta = \frac{\log N}{2(\sigma_1-\sigma_2)N}$ 可知当 N 趋于无穷时有 $\eta\to 0$ 以及 $N\eta\to+\infty$，故可得 $\eta\hat{o}(1) = o(\eta)$。由 (C-37) 及 $\eta = o(N\eta^2)$ 可知

$$\begin{aligned}\frac{\pi_N^{(1)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle} &= 1+\eta\cdot\frac{\alpha}{\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}+o(\eta)\\ &= 1+o(N\eta^2).\end{aligned}$$

综合这两项的结论，可知对充分大的 N 有

$$\begin{aligned}\frac{(1+\eta\sigma_1)^{2N}\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}{\pi_N^{(1)}(\boldsymbol{\phi}_0)} &= \left[\frac{(\lambda_{11}(\mathbf{G}))^N}{(1+\eta\sigma_1)^{2N}}\cdot\frac{\pi_N^{(1)}(\boldsymbol{\phi}_0)}{(\lambda_{11}(\mathbf{G}))^N\cdot\langle\mathbf{e}_{11},\boldsymbol{\theta}_0\rangle}\right]^{-1}\\ &= \left[(1+N\eta^2\tau_1+o(N\eta^2))(1+o(N\eta^2))\right]^{-1}\\ &= 1-N\eta^2\tau_1+o(N\eta^2)\\ &> 1-2N\eta^2\tau_1.\end{aligned}$$

从而得

$$\mathbb{E}\left[\nu(\boldsymbol{\phi}_N)|\boldsymbol{\phi}_0\right] > (1-2\tau_1 N\eta^2)\bar{v}_N(\boldsymbol{\phi}_0) = 1-C_0\cdot\frac{\log^2 N}{N},$$

故

$$\mathbb{E}\left[\nu^{\mathsf{C}}(\boldsymbol{\phi}_N)|\boldsymbol{\phi}_0\right] < C_0\cdot\frac{\log^2 N}{N}.$$

利用 Markov 不等式即得 (5-42) 的结论。最后，(5-43) 中关于 $\rho^{\mathsf{C}}(\boldsymbol{\phi}_N)$ 的结论可立即由不等式 $\rho^{\mathsf{C}}(\boldsymbol{\phi}_N)\leqslant\sigma_1\nu^{\mathsf{C}}(\boldsymbol{\phi}_N)$ 推出。 □

附录 D　第 6 章中的证明

D.1　命题 6.1的证明

对任意 $y' \in \mathcal{Y}$，可得

$$\frac{\partial}{\partial b(y')} P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x) = P_Y(y')\delta_{y,y'},$$

以及

$$\nabla_{\mathbf{g}(y')} P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x) = P_Y(y)\mathbf{f}(x)\delta_{y,y'},$$

其中 $\delta_{y,y'}$ 表示 Kronecker delta 符号。由此可知

$$\begin{aligned}
\frac{\partial}{\partial b(y')} L(\mathbf{f},\mathbf{g},b) &= -2 \sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} \left[\frac{P_{Y|X}(y|x) - P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)}{P_Y(y)} \cdot \frac{\partial}{\partial b(y')} P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x) \right] \\
&= -2 \sum_{x\in\mathcal{X}} P_X(x) \left[P_{Y|X}(y'|x) - P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y'|x) \right] \\
&= -2 \sum_{x\in\mathcal{X}} P_X(x) \left[P_{Y|X}(y'|x) - P_Y(y')(1 + \mathbf{f}(x)\mathbf{g}(y') + b(y')) \right] \\
&= -2P_Y(y') \left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}} \mathbf{g}(y') + b(y') \right]
\end{aligned}$$

及

$$\begin{aligned}
\nabla_{\mathbf{g}(y')} L(\mathbf{f},\mathbf{g},b) &= -2 \sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} \left[\frac{P_{Y|X}(y|x) - P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)}{P_Y(y)} \cdot \nabla_{\mathbf{g}(y')} P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x) \right] \\
&= -2 \sum_{x\in\mathcal{X}} P_X(x) \left[P_{Y|X}(y'|x) - P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y'|x) \right] \mathbf{f}(x) \\
&= -2P_Y(y') \sum_{x\in\mathcal{X}} \left[P_{X|Y}(x|y') - P_X(x)(1 + \mathbf{f}^{\mathsf{T}}(x)\mathbf{g}(y') + b(y')) \right] \mathbf{f}(x) \\
&= -2P_Y(y') \left(\mathbb{E}[\mathbf{f}(X)|Y = y'] - \boldsymbol{\mu}_{\mathbf{f}} - \mathbb{E}[\mathbf{f}(X)\mathbf{f}^{\mathsf{T}}(X)]\mathbf{g}(y') - \boldsymbol{\mu}_{\mathbf{f}} b(y') \right) \\
&= -2P_Y(y') \left(\mathbb{E}\left[\tilde{\mathbf{f}}(X) \middle| Y = y' \right] - \mathbb{E}\left[\mathbf{f}(X)\mathbf{f}^{\mathsf{T}}(X) \right] \mathbf{g}(y') - \boldsymbol{\mu}_{\mathbf{f}} b(y') \right).
\end{aligned}$$

又因最优参数 $\mathbf{g}^*$ 与 b^* 满足对任意 $y \in \mathcal{Y}$，

$$\frac{\partial}{\partial b(y)} = 0, \quad \nabla_{\mathbf{g}(y)} L(\mathbf{f},\mathbf{g},b) = \mathbf{0},$$

故

$$\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}^{*}(y)+b^{*}(y)=\mathbf{0} \tag{D-1}$$

$$\mathbb{E}\big[\mathbf{f}(X)\mathbf{f}^{\mathsf{T}}(X)\big]\mathbf{g}^{*}(y)+\boldsymbol{\mu}_{\mathbf{f}}b^{*}(y)=\mathbb{E}\big[\tilde{\mathbf{f}}(X)\big|Y=y\big]. \tag{D-2}$$

将 (D-1) 代入 (D-2)，得

$$\mathbb{E}\big[\mathbf{f}(X)\mathbf{f}^{\mathsf{T}}(X)\big]\mathbf{g}^{*}(y)-\boldsymbol{\mu}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}^{*}(y)=\boldsymbol{\Lambda}_{\mathbf{f}}\mathbf{g}^{*}(y)=\mathbb{E}\big[\tilde{\mathbf{f}}(X)\big|Y=y\big],$$

从而得 (6-3a)。最后，可立即由 (D-1) 推出 (6-3b)。

D.2　命题 6.2 的证明

首先给出如下引理。

引理 D.1： 对任意的 $\mathbf{f}$、$\mathbf{g}$ 及 b，有

$$L(\mathbf{f},\mathbf{g},b)=\chi^{2}(P_{XY},P_{X}P_{Y})-2H(\mathbf{f},\mathbf{g})+\boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}}+\sum_{y\in\mathcal{Y}}P_{Y}(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y)+b(y)\right]^{2},$$

其中 $\chi^{2}(P_{XY},P_{X}P_{Y})$ 为联合分布 P_{XY} 到乘积分布 $P_{X}P_{Y}$ 的 χ^{2} 散度：

$$\chi^{2}(P_{XY},P_{X}P_{Y})=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\big[P_{XY}(x,y)-P_{X}(x)P_{Y}(y)\big]^{2}}{P_{X}(x)P_{Y}(y)}.$$

基于引理 D.1，命题 6.2 可证明如下。注意到对所有 $\boldsymbol{\theta}$、$\mathbf{g}$ 及 b，有

$$\begin{aligned}
L(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g},b)&=\chi^{2}(P_{XY},P_{X}P_{Y})-2H(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g})+\boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}}+\sum_{y\in\mathcal{Y}}P_{Y}(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y)+b(y)\right]^{2}\\
&\geqslant\chi^{2}(P_{XY},P_{X}P_{Y})-2H(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g}) &\text{(D-3)}\\
&\geqslant\chi^{2}(P_{XY},P_{X}P_{Y})-2H(\mathbf{f}_{\boldsymbol{\theta}_{\mathrm{H}}},\mathbf{g}_{\mathrm{H}}) &\text{(D-4)}\\
&=\chi^{2}(P_{XY},P_{X}P_{Y})-2H(\mathbf{f}_{\boldsymbol{\theta}_{\mathrm{H}}},\tilde{\mathbf{g}}_{\mathrm{H}}) &\text{(D-5)}\\
&=L(\mathbf{f}_{\boldsymbol{\theta}_{\mathrm{H}}},\tilde{\mathbf{g}}_{\mathrm{H}},b^{*}), &\text{(D-6)}
\end{aligned}$$

其中 (D-3) 基于

$$\boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}}+\sum_{y\in\mathcal{Y}}P_{Y}(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y)+b(y)\right]^{2}\geqslant 0. \tag{D-7}$$

此外，式 (D-4)成立因为 $\boldsymbol{\theta}_{\mathrm{H}}$ 与 $\mathbf{g}_{\mathrm{H}}$ 为最大化 $H(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{g})$ 的参数；式 (D-5) 依据 $H(\mathbf{f}_{\boldsymbol{\theta}_{\mathrm{H}}}, \mathbf{g}_{\mathrm{H}}) = H(\mathbf{f}_{\boldsymbol{\theta}_{\mathrm{H}}}, \tilde{\mathbf{g}}_{\mathrm{H}})$；式 (D-6) 的依据是，(D-7) 的等号对 $\boldsymbol{\theta}_{\mathrm{H}}$, $\tilde{\mathbf{g}}_{\mathrm{H}}$ 及 $b^*(y) = -\tilde{\mathbf{g}}_{\mathrm{H}}^{\mathsf{T}}(y)\mathbb{E}[\mathbf{f}(X;\boldsymbol{\theta}_{\mathrm{H}})]$ 成立。

故可得 $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{\mathrm{H}}$，$\mathbf{g}^* = \tilde{\mathbf{g}}_{\mathrm{H}}^*$。且对任意 $y \in \mathcal{Y}$，有

$$b^*(y) = -\tilde{\mathbf{g}}_{\mathrm{H}}^{\mathsf{T}}(y)\mathbb{E}[\mathbf{f}(X;\boldsymbol{\theta}_{\mathrm{H}})] = -\boldsymbol{\mu}_{\mathbf{f}^*}^{\mathsf{T}}\mathbf{g}^*(y)$$

其中 $\boldsymbol{\mu}_{\mathbf{f}^*} \triangleq \mathbb{E}[\mathbf{f}(X;\boldsymbol{\theta}^*)] = \mathbb{E}[\mathbf{f}(X;\boldsymbol{\theta}_{\mathrm{H}})]$。

现在只需证明 引理 D.1。

证明 (引理 D.1 的证明) 注意到

$$\begin{aligned}
L(\mathbf{f},\mathbf{g},b) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{\left[P_{XY}(x,y) - P_X(x)P_{Y|X}^{(\mathbf{f},\mathbf{g},b)}(y|x)\right]^2}{P_X(x)P_Y(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{1}{P_X(x)P_Y(y)} \cdot \left[P_{XY}(x,y) - P_X(x)P_Y(y)\left(1 + \mathbf{f}^{\mathsf{T}}(x)\mathbf{g}(y) + b(y)\right)\right]^2 \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{\left[P_{XY}(x,y) - P_X(x)P_Y(y)\right]^2}{P_X(x)P_Y(y)} \\
&\quad - 2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left([P_{XY}(x,y) - P_X(x)P_Y(y)] \cdot [\mathbf{f}^{\mathsf{T}}(x)\mathbf{g}(y) + b(y)]\right) \\
&\quad + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_Y(y)\left(\mathbf{f}^{\mathsf{T}}(x)\mathbf{g}(y) + b(y)\right)^2 \\
&= \chi^2(P_{XY}, P_XP_Y) - 2\,\mathbb{E}\left[\tilde{\mathbf{f}}^{\mathsf{T}}(X)\tilde{\mathbf{g}}(Y)\right] \\
&\quad + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_Y(y) \cdot \left[\tilde{\mathbf{f}}^{\mathsf{T}}(x)\tilde{\mathbf{g}}(y) + \tilde{\mathbf{f}}^{\mathsf{T}}(x)\boldsymbol{\mu}_{\mathbf{g}} + \left(\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right)\right]^2 \\
&= \chi^2(P_{XY}, P_XP_Y) - 2\,\mathbb{E}\left[\tilde{\mathbf{f}}^{\mathsf{T}}(X)\tilde{\mathbf{g}}(Y)\right] + \mathrm{tr}\left\{\boldsymbol{\Lambda}_{\tilde{\mathbf{f}}}\boldsymbol{\Lambda}_{\tilde{\mathbf{g}}}\right\} \\
&\quad + \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}} + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_Y(y)\left(\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right)^2 \\
&= \chi^2(P_{XY}, P_XP_Y) - 2H(\mathbf{f},\mathbf{g}) + \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\tilde{\mathbf{f}}}\boldsymbol{\mu}_{\mathbf{g}} + \sum_{y\in\mathcal{Y}} P_Y(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right]^2,
\end{aligned}$$

其中倒数第二个等式依据

$$\begin{aligned}
\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_Y(y) \cdot \left[\tilde{\mathbf{f}}^{\mathsf{T}}(x)\tilde{\mathbf{g}}(y)\right]^2 &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_Y(y) \cdot \mathrm{tr}\left(\tilde{\mathbf{f}}(x)\tilde{\mathbf{f}}^{\mathsf{T}}(x)\tilde{\mathbf{g}}(y)\tilde{\mathbf{g}}^{\mathsf{T}}(y)\right) \\
&= \mathrm{tr}\left\{\sum_{x\in\mathcal{X}} P_X(x)\tilde{\mathbf{f}}(x)\tilde{\mathbf{f}}^{\mathsf{T}}(x)\sum_{y\in\mathcal{Y}} P_Y(y)\tilde{\mathbf{g}}(y)\tilde{\mathbf{g}}^{\mathsf{T}}(y)\right\}
\end{aligned}$$

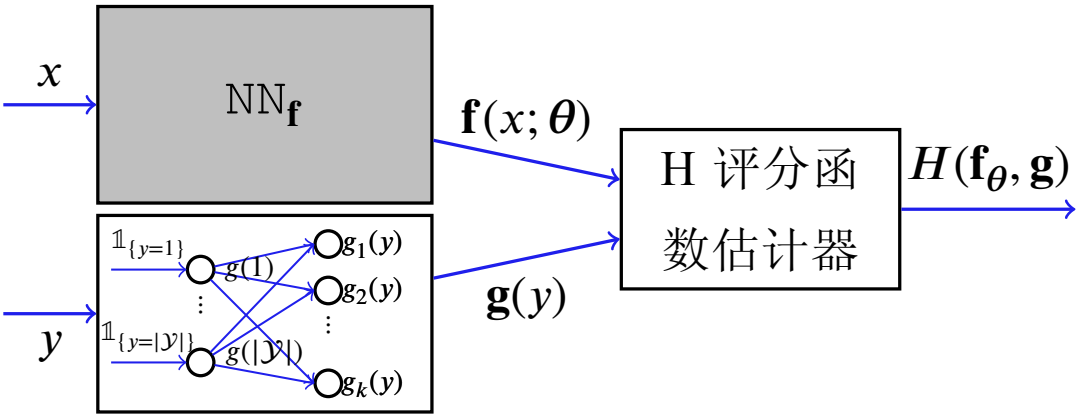


图 D.1　联合训练最大相关回归参数 $\boldsymbol{\theta}$ 及 $\mathbf{g}$ 的网络架构，其中：$\mathrm{NN}_{\mathbf{f}}$ 用于从输入 x 中生成特征 $\mathbf{f}(x;\boldsymbol{\theta})$，且 $\boldsymbol{\theta}$ 表示 $\mathrm{NN}_{\mathbf{f}}$ 的参数；$\mathbf{g}$ 可由一个全连接层生成。

$$= \operatorname{tr}\left\{\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\Lambda}_{\mathbf{g}}\right\}$$

以及

$$\begin{aligned}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_X(x)P_Y(y)\left[\tilde{\mathbf{f}}^{\mathsf{T}}(x)\boldsymbol{\mu}_{\mathbf{g}}\right]^2 &= \sum_{x\in\mathcal{X}} P_X(x)\boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\tilde{\mathbf{f}}(x)\tilde{\mathbf{f}}^{\mathsf{T}}(x)\boldsymbol{\mu}_{\mathbf{g}} \\ &= \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\left(\sum_{x\in\mathcal{X}} P_X(x)\tilde{\mathbf{f}}(x)\tilde{\mathbf{f}}^{\mathsf{T}}(x)\right)\boldsymbol{\mu}_{\mathbf{g}} \\ &= \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}}.\end{aligned}$$

□

D.3　优化 H 评分函数的神经网络架构

图 D.1 给出了联合优化 $\boldsymbol{\theta}$ 及 $\mathbf{g}$ 的神经网络架构。给定数据与标签的样本 (x, y)，$\mathbf{f}(x;\boldsymbol{\theta})$ 由网络 $\mathrm{NN}_{\mathbf{f}}$ 生成，其中 $\boldsymbol{\theta}$ 对应 $\mathrm{NN}_{\mathbf{f}}$ 中所有可训练参数。此外因 Y 离散，$\mathbf{g}(y)$ 可由输入为 y 的单点编码 $[\mathbb{1}_{\{y=1\}},\cdots,\mathbb{1}_{\{y=|\mathcal{Y}|\}}]^{\mathsf{T}}$、含 k 个输出节点的单个全连接层生成，且该全连接层权重对应于 $\mathbf{g}(1),\cdots,\mathbf{g}(|\mathcal{Y}|)$。

在此基础上，将损失函数取为 $-H(\mathbf{f}_{\boldsymbol{\theta}},\mathbf{g})$，则训练该网络可得最优参数 $\boldsymbol{\theta}_{\mathrm{H}}$ 及 $\mathbf{g}_{\mathrm{H}}$。

D.4　命题 6.3 的证明

由 CDM 定义 (2-3) 可知

$$\chi^2(P_{XY}, P_X P_Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{\left[P_{XY}(x,y) - P_X(x)P_Y(y)\right]^2}{P_X(x)P_Y(y)} = \|\tilde{\mathbf{B}}\|_{\mathrm{F}}^2,$$

此外，注意到 H 评分函数可等价表示为 (可参考第 3.5 节中的讨论)

$$H(\mathbf{f},\mathbf{g}) = \frac{1}{2}\left[\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - \left\|\tilde{\mathbf{B}} - \boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2\right],$$

故由 引理 D.1 可得

$$\begin{aligned} L(\mathbf{f},\mathbf{g},b) &= \chi^2(P_{XY}, P_X P_Y) - 2H(\mathbf{f},\mathbf{g}) + \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}} + \sum_{y\in\mathcal{Y}} P_Y(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right]^2, \\ &= \left\|\tilde{\mathbf{B}} - \boldsymbol{\Xi}^Y\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2 + \boldsymbol{\mu}_{\mathbf{g}}^{\mathsf{T}}\boldsymbol{\Lambda}_{\mathbf{f}}\boldsymbol{\mu}_{\mathbf{g}} + \sum_{y\in\mathcal{Y}} P_Y(y)\left[\boldsymbol{\mu}_{\mathbf{f}}^{\mathsf{T}}\mathbf{g}(y) + b(y)\right]^2. \end{aligned}$$

D.5　定理 6.1 的证明

基于真实后验概率分布 $P_{Y|X}$ 给出的最大后验概率预测准确率 ACC* 可表示为

$$\begin{aligned} \mathrm{ACC}^* &= \sum_{x\in\mathcal{X}} P_X(x) \max_{y\in\mathcal{Y}} P_{Y|X}(y|x) \\ &= \sum_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} P_{XY}(x,y). \end{aligned}$$

令 $\mathbf{B}\in\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ 表示为 X 与 Y 之间的散度转移矩阵 (参见定义 2.3)，则有

$$\begin{aligned} \|\mathbf{B}\|_{\mathsf{F}}^2 = \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 + 1 &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{P_{XY}(x,y)^2}{P_X(x)P_Y(y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \frac{1}{P_Y(y)} \cdot P_{XY}(x,y) \cdot P_{Y|X}(y|x) \\ &\leqslant \frac{1}{p_{\min}} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y) \cdot P_{Y|X}(y|x) \\ &\leqslant \frac{1}{p_{\min}} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \left(\max_{y'\in\mathcal{Y}} P_{XY}(x,y')\right) \cdot P_{Y|X}(y|x) \\ &= \frac{1}{p_{\min}} \sum_{x\in\mathcal{X}} \left(\max_{y'\in\mathcal{Y}} P_{XY}(x,y')\right) \cdot \left(\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\right) \\ &= \frac{\mathrm{ACC}^*}{p_{\min}}, \end{aligned}$$

从而 (6-12) 中的第一个不等号成立。第二个不等号可立即由

$$H_Y(\mathbf{f}) \leqslant \frac{1}{2}\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2.$$

推出。为证明 (6-13)，对给定观测 x，令 $y^*(x)$ 与 $\hat{y}(x)$ 分别表示基于真实后验概率分布及估计所得 $P_{Y|X}^{(\mathbf{f},\mathbf{g}^*,b^*)}(y|x)$ 的最大后验概率预测结果，即

$$y^*(x) \triangleq \arg\max_{y\in\mathcal{Y}} P_{Y|X}(y|x),$$
$$\hat{y}(x) \triangleq \arg\max_{y\in\mathcal{Y}} P_{Y|X}^{(\mathbf{f},\mathbf{g}^*,b^*)}(y|x).$$

此外，定义集合 $\mathcal{X}' \subset \mathcal{X}$ 为

$$\mathcal{X}' \triangleq \left\{x \in \mathcal{X} :\ y^*(x) \neq \hat{y}(x)\right\},$$

并令

$$\hat{P}_{XY}(x,y) \triangleq P_X(x) P_{Y|X}^{(\mathbf{f},\mathbf{g}^*,b^*)}(y|x),$$

则可得

$$\begin{aligned}
&\mathrm{ACC}^* - \mathrm{ACC} \\
&\quad= \sum_{x\in\mathcal{X}} P_{XY}(x,y^*(x)) - \sum_{x\in\mathcal{X}} P_{XY}(x,\hat{y}(x)) \\
&\quad= \sum_{x\in\mathcal{X}'} \left[P_{XY}(x,y^*(x)) - P_{XY}(x,\hat{y}(x))\right] \\
&\quad\leqslant \sum_{x\in\mathcal{X}'} \left[P_{XY}(x,y^*(x)) - P_{XY}(x,\hat{y}(x)) + \hat{P}_{XY}(x,\hat{y}(x)) - \hat{P}_{XY}(x,y^*(x))\right] \\
&\quad\leqslant \sum_{x\in\mathcal{X}'} \Big(\left|P_{XY}(x,y^*(x)) - \hat{P}_{XY}(x,y^*(x))\right| + \left|P_{XY}(x,\hat{y}(x)) - \hat{P}_{XY}(x,\hat{y}(x))\right|\Big),
\end{aligned}$$

其中第一个不等号的依据为

$$\begin{aligned}
\hat{P}_{XY}(x,\hat{y}(x)) &= P_X(x) P_{Y|X}^{(\mathbf{f},\mathbf{g}^*,b^*)}(\hat{y}(x)|x) \\
&\geqslant P_X(x) P_{Y|X}^{(\mathbf{f},\mathbf{g}^*,b^*)}(y^*(x)|x) \\
&= \hat{P}_{XY}(x,y^*(x)).
\end{aligned}$$

在此基础上，令 $\boldsymbol{\Xi}^Y$ 与 $\boldsymbol{\Xi}^X$ 分别表示 $\tilde{\mathbf{f}}$ 与 $\mathbf{g}^*$ 由 (6-8) 定义的矩阵表示，则

$$\begin{aligned}
\left\|\tilde{\mathbf{B}} - \boldsymbol{\Xi}^X \left(\boldsymbol{\Xi}^Y\right)^{\mathsf{T}}\right\|_{\mathrm{F}}^2 &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{1}{P_X(x)P_Y(y)} \Big[P_{XY}(x,y) - P_X(x)P_Y(y) \\
&\qquad - P_X(x)P_Y(y)\tilde{\mathbf{f}}^{\mathsf{T}}(x)\mathbf{g}^*(y)\Big]^2 \qquad \text{(D-8)} \\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \frac{\left[P_{XY}(x,y) - \hat{P}_{XY}(x,y)\right]^2}{P_X(x)P_Y(y)} \qquad \text{(D-9)}
\end{aligned}$$

$$\geqslant \frac{1}{p_{\max}} \sum_{x\in\mathcal{X}} \frac{1}{P_X(x)} \sum_{y\in\mathcal{Y}} \left[P_{XY}(x,y) - \hat{P}_{XY}(x,y)\right]^2 \tag{D-10}$$

$$\geqslant \frac{1}{p_{\max}} \sum_{x\in\mathcal{X}'} \frac{1}{P_X(x)} \sum_{y\in\mathcal{Y}} \left[P_{XY}(x,y) - \hat{P}_{XY}(x,y)\right]^2 \tag{D-11}$$

$$\begin{aligned}\geqslant \frac{1}{p_{\max}} \sum_{x\in\mathcal{X}'} \frac{1}{P_X(x)} \Big(&\left[P_{XY}(x,y^*(x)) - \hat{P}_{XY}(x,y^*(x))\right]^2 \\ &+ \left[P_{XY}(x,\hat{y}(x)) - \hat{P}_{XY}(x,\hat{y}(x))\right]^2 \Big)\end{aligned} \tag{D-12}$$

$$\begin{aligned}\geqslant \frac{1}{2p_{\max}} \sum_{x\in\mathcal{X}'} \frac{1}{P_X(x)} \Big(&\left|P_{XY}(x,y^*(x)) - \hat{P}_{XY}(x,y^*(x))\right| \\ &+ \left|P_{XY}(x,\hat{y}(x)) - \hat{P}_{XY}(x,\hat{y}(x))\right| \Big)^2\end{aligned} \tag{D-13}$$

$$\begin{aligned}\geqslant \frac{1}{2p_{\max}} \Bigg[\sum_{x\in\mathcal{X}'} \Big(&\left|P_{XY}(x,y^*(x)) - \hat{P}_{XY}(x,y^*(x))\right| \Big) \\ &+ \left|P_{XY}(x,\hat{y}(x)) - \hat{P}_{XY}(x,\hat{y}(x))\right| \Big) \Bigg]^2 \cdot \left(\sum_{x\in\mathcal{X}'} P_X(x) \right)^{-1}\end{aligned} \tag{D-14}$$

$$= \frac{1}{2p_{\max}} (\mathrm{ACC}^* - \mathrm{ACC})^2 \cdot \left(\sum_{x\in\mathcal{X}'} P_X(x) \right)^{-1} \tag{D-15}$$

$$\geqslant \frac{1}{2p_{\max}} (\mathrm{ACC}^* - \mathrm{ACC})^2, \tag{D-16}$$

其中 (D-9) 依据 $\hat{P}_{XY}(x,y) = P_X(x)P_Y(y)(1+\tilde{\mathbf{f}}^{\mathsf{T}}(x)\mathbf{g}^*))$，(D-13) 基于不等式 $a^2+b^2 \geqslant \frac{(a+b)^2}{2}$，(D-14) 基于 Cauchy–Schwarz 不等式。

最后，因 (6-3a) 等价于

$$\boldsymbol{\Xi}^Y = \mathbf{B}\boldsymbol{\Xi}^X \left(\left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}} \boldsymbol{\Xi}^X \right)^{-1}, \tag{D-17}$$

从而推出

$$\left\| \tilde{\mathbf{B}} - \boldsymbol{\Xi}^Y \left(\boldsymbol{\Xi}^X\right)^{\mathsf{T}} \right\|_{\mathrm{F}}^2 = \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - 2H_Y(\mathbf{f}),$$

因此得

$$(\mathrm{ACC}^* - \mathrm{ACC})^2 \leqslant 2 \left[\left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - 2H_Y(\mathbf{f}) \right] p_{\max}.$$

D.6　命题 6.4 的证明

充分性可由立即由

$$H_Y(\mathbf{f}(X)) = H_Y\left(\boldsymbol{\Lambda}^{-1/2}\tilde{\mathbf{f}}(X)\right)$$

$$= \mathbb{E}\left[\left\|\mathbb{E}\left[\boldsymbol{\Lambda}^{-1/2}\tilde{\mathbf{f}}(X)\middle|Y\right]\right\|^2\right]$$

推出。

下面证明必要性。为证明 (a)，首先注意到

$$\boldsymbol{\Lambda}_{\mathbf{Af}(X)+\mathbf{a}} = \mathbf{A}\boldsymbol{\Lambda}_{\mathbf{f}}\mathbf{A}^{\mathsf{T}}$$

及

$$\operatorname{cov}\left(\mathbb{E}[\mathbf{Af}(X)+\mathbf{a}|Y]\right) = \mathbf{A}\operatorname{cov}\left(\mathbb{E}[\mathbf{f}(X)|Y]\right)\mathbf{A}^{\mathsf{T}}.$$

因此，根据 (6-11) 中的第二个等式可得

$$H_Y(\mathbf{Af}(X)+\mathbf{a}) = H_Y(\mathbf{f}(X)).$$

为证明 (b)，注意到 (6-16a) 可立即由 (6-11) 推出。为验证 (6-16b)，根据全协方差公式可知

$$\mathbf{I} = \operatorname{cov}(\mathbb{E}[\mathbf{f}(X)|Y]) + \mathbb{E}[\operatorname{cov}(\mathbf{f}(X)|Y)],$$

从而由 (6-15) 可知

$$\begin{aligned} H_Y(\mathbf{f}(X)) &= \operatorname{tr}\left\{\operatorname{cov}(\mathbb{E}[\mathbf{f}(X)|Y])\right\} \\ &= \operatorname{tr}\left\{\mathbf{I} - \mathbb{E}[\operatorname{cov}(\mathbf{f}(X)|Y)]\right\} \\ &= k - \operatorname{tr}\left\{\mathbb{E}[\operatorname{cov}(\mathbf{f}(X)|Y)]\right\} \\ &= k - \mathbb{E}[\operatorname{tr}\left\{\operatorname{cov}(\mathbf{f}(X)|Y)\right\}] \\ &= k - \mathbb{E}\left[\|\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y]\|^2\right], \end{aligned}$$

其中最后的等号成立依据

$$\begin{aligned} \operatorname{tr}\left\{\operatorname{cov}(\mathbf{f}(X)|Y)\right\} &= \operatorname{tr}\left\{\mathbb{E}\left[(\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y])(\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y])^{\mathsf{T}}\middle|Y\right]\right\} \\ &= \mathbb{E}\left[\operatorname{tr}\left\{(\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y])(\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y])^{\mathsf{T}}\right\}\middle|Y\right] \\ &= \mathbb{E}\left[\|\mathbf{f}(X) - \mathbb{E}[\mathbf{f}(X)|Y]\|^2\middle|Y\right]. \end{aligned}$$

D.7　引理 6.1 的证明

为证明第一条性质，对 $i=1,\cdots,d$，定义 $|\mathcal{X}_i|\times(|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdots|\mathcal{X}_d|)$ 矩阵 $\mathbf{B}_i$ 为

$$B_i(x_i';x^d)=\begin{cases}\dfrac{\sqrt{P_{X^d}(x^d)}}{\sqrt{P_{X_i}(x_i')}} & \text{若 } x_i'=x_i\\ 0 & \text{其他情况。}\end{cases} \tag{D-18}$$

在此基础上，定义 $(|\mathcal{X}_1|+\cdots+|\mathcal{X}_m|)\times(|\mathcal{X}_1|\cdots|\mathcal{X}_m|)$ 维矩阵

$$\mathbf{B}_0\triangleq\begin{bmatrix}\mathbf{B}_1\\ \vdots\\ \mathbf{B}_d\end{bmatrix}, \tag{D-19}$$

则可验证

$$\mathbf{B}=\mathbf{B}_0\mathbf{B}_0^{\mathsf{T}}, \tag{D-20}$$

从而可知 $\mathbf{B}$ 半正定。

为导出第二条性质，注意到 $\boldsymbol{\psi}^{(0)}$ 为矩阵 $\mathbf{B}$ 对应于特征值 d 的特征向量，故 $\left(\boldsymbol{\psi}^{(0)}\right)^{\mathsf{T}}\mathbf{B}\boldsymbol{\psi}^{(0)}=d$. 此外，由[35] 知 $\mathbf{B}_{ij}$ 的最大奇异值为 1，亦即 $\left\|\mathbf{B}_{ij}\right\|_{\mathrm{s}}=1$，其中 $\|\cdot\|_{\mathrm{s}}$ 表示矩阵的谱范数。

因此，对于由诸 $|\mathcal{X}_i|$ 维向量 $\boldsymbol{\psi}_i$ 构成的向量 $\boldsymbol{\psi}=\left[\boldsymbol{\psi}_1^{\mathsf{T}},\cdots,\boldsymbol{\psi}_d^{\mathsf{T}}\right]^{\mathsf{T}}$ 有

$$\begin{aligned}\boldsymbol{\psi}^{\mathsf{T}}\mathbf{B}\boldsymbol{\psi}=\sum_{i=1}^{d}\sum_{j=1}^{d}\boldsymbol{\psi}_i^{\mathsf{T}}\mathbf{B}_{ij}\boldsymbol{\psi}_j &\leqslant\sum_{i=1}^{d}\sum_{j=1}^{d}\|\boldsymbol{\psi}_i\|\cdot\left\|\mathbf{B}_{ij}\right\|_{\mathrm{s}}\cdot\|\boldsymbol{\psi}_j\|\\ &=\left(\sum_{i=1}^{d}\|\boldsymbol{\psi}_i\|\right)^2\\ &\leqslant d\sum_{i=1}^{d}\|\boldsymbol{\psi}_i\|^2=d\|\boldsymbol{\psi}\|^2,\end{aligned}$$

其中第二个不等号可由算术平均不超过平方平均推出。从而可得

$$\max_{\boldsymbol{\psi}:\ \|\boldsymbol{\psi}\|=1}\boldsymbol{\psi}^{\mathsf{T}}\mathbf{B}\boldsymbol{\psi}=d,$$

即 $\mathbf{B}$ 最大特征值为等于 d。

为验证第三条性质，构造

$$\boldsymbol{\psi}' = \begin{bmatrix} \boldsymbol{\psi}_1' \\ \mathbf{0} \end{bmatrix},$$

其中 $\boldsymbol{\psi}_1' \in \mathbb{R}^{|\mathcal{X}_1|}$ 满足 $\langle \boldsymbol{\psi}_1', \mathbf{v}_1 \rangle = 0$ 以及 $\|\boldsymbol{\psi}_1'\|^2 = 1$，且式中 $\mathbf{0}$ 为 $(m - |\mathcal{X}_1|)$ 维的零向量。则可验证 $\left\langle \boldsymbol{\psi}', \boldsymbol{\psi}^{(0)} \right\rangle = 0$ 及 $\left\|\boldsymbol{\psi}'\right\|^2 = 1$. 此外，注意到 $\mathbf{B}$ 第二大特征值 $\lambda^{(1)}$ 可表示为

$$\lambda^{(1)} = \max_{\boldsymbol{\psi}\,:\,\|\boldsymbol{\psi}\|=1, \left\langle \boldsymbol{\psi}, \boldsymbol{\psi}^{(0)} \right\rangle = 0} \boldsymbol{\psi}^{\mathsf{T}} \mathbf{B} \boldsymbol{\psi},$$

由此推出 $\lambda^{(1)} \geqslant \left(\boldsymbol{\psi}'\right)^{\mathsf{T}} \mathbf{B} \boldsymbol{\psi}' = \|\boldsymbol{\psi}_1'\|^2 = 1.$

为检验第四条性质, 定义 $(d-1)$ 维子空间

$$\mathcal{S}_{\text{eig}} \triangleq \left\{ \boldsymbol{\psi} = \left[\alpha_1 \mathbf{v}_1^{\mathsf{T}}, \cdots, \alpha_d \mathbf{v}_d^{\mathsf{T}} \right]^{\mathsf{T}} : \sum_{i=1}^{d} \alpha_i = 0 \right\}, \tag{D-21}$$

则对任意 $\boldsymbol{\psi} \in \mathcal{S}_{\text{eig}}$，由 $\mathbf{B}_{ij} \mathbf{v}_j = \mathbf{v}_i$ 可得 $\mathbf{B}\boldsymbol{\psi} = \mathbf{0}_m$，其中 $\mathbf{0}_m$ 为 $\mathbb{R}^m$ 中的零向量。因此, $\mathcal{S}_{\text{eig}}$ 为 $\mathbf{B}$ 对应于 $d-1$ 个零特征值的特征空间。根据 $\mathbf{B}$ 的半正定性可不妨假设 $\mathcal{S}_{\text{eig}}$ 由 $\boldsymbol{\psi}^{(m-d+1)}, \cdots, \boldsymbol{\psi}^{(m-1)}$ 张成，对应特征值 $\lambda^{(m-d+1)} = \cdots = \lambda^{(m-1)} = 0$。

为证明最后一条性质，对任意 $\ell = 1, \cdots, m-d$，由 $\left\langle \boldsymbol{\psi}^{(\ell)}, \boldsymbol{\psi}^{(0)} \right\rangle = 0$ 可得

$$\sum_{i=1}^{d} \left\langle \boldsymbol{\psi}_i^{(\ell)}, \mathbf{v}_i \right\rangle = 0,$$

故由第三条性质可得

$$\boldsymbol{\psi}' = \begin{bmatrix} \left\langle \boldsymbol{\psi}_1^{(\ell)}, \mathbf{v}_1 \right\rangle \mathbf{v}_1 \\ \vdots \\ \left\langle \boldsymbol{\psi}_d^{(\ell)}, \mathbf{v}_d \right\rangle \mathbf{v}_d \end{bmatrix} \in \mathcal{S}_{\text{eig}}.$$

因此有 $\langle \boldsymbol{\psi}', \boldsymbol{\psi}^{(\ell)} \rangle = 0$，亦即

$$\sum_{i=1}^{d} \left\langle \boldsymbol{\psi}_i^{(\ell)}, \mathbf{v}_i \right\rangle^2 = 0,$$

从而可推出 $\left\langle \boldsymbol{\psi}_i^{(\ell)}, \mathbf{v}_i \right\rangle = 0, \quad i = 1, \cdots, d.$

D.8　定理 6.2 的证明

首先，使用 $\frac{1}{2}\epsilon^2$ 替代 δ 以便于局部信息几何方法中的表述。则约束 (6-17) 可表示为

$$I(U;X^d) \leqslant \frac{1}{2}\epsilon^2, \tag{D-22}$$

其中 ϵ 为一小量。根据 (6-18) 及 (D-22) 可知对所有的 u，条件分布 $P_{X^d|U=u}$ 可表达为边缘分布的扰动：

$$P_{X^d|U}(x^d|u) = P_{X^d}(x^d) + \epsilon\sqrt{P_{X^d}(x^d)}\phi_u(x^d) \tag{D-23}$$

其中 ϕ_u 可看作是 $(|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdots|\mathcal{X}_d|)$ 维的向量。此外，由 K-L 散度的二阶 Taylor 展开 (参考命题 2.1) 可知

$$\begin{aligned} I(U;X^d) &= \mathbb{E}_U\left[D(P_{X^d|U}\|P_{X^d})\right] \\ &= \frac{1}{2}\epsilon^2\mathbb{E}_U\left[\left\|\boldsymbol{\phi}_U\right\|^2\right] + o(\epsilon^2), \end{aligned}$$

其中 $\|\cdot\|$ 表示 l_2 范数。由于 ϵ 很小，忽略高阶项之后可将约束 $I(U;X^d) \leqslant \frac{1}{2}\epsilon^2$ 化简为

$$\mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2] \leqslant 1. \tag{D-24}$$

进一步，注意到目标函数 $\ell(X^d|U)$ 亦可表达为互信息的形式：

$$D(P_{X^d}\|P_{X_1}\cdots P_{X_d}) - D(P_{X^d}\|P_{X_1}\cdots P_{X_d}|U) = \sum_{i=1}^{d} I(U;X_i) - I(U;X^d). \tag{D-25}$$

对任意 i，互信息 $I(U;X_i)$ 可由相应的 l_2 范数近似：

$$I(U;X_i) = \frac{1}{2}\epsilon^2\mathbb{E}_U[\|\boldsymbol{\psi}_{i,U}\|^2] + o(\epsilon^2),$$

其中当 $U=u$ 时，对应 $\boldsymbol{\psi}_{i,u}$ 定义为相应的 $|\mathcal{X}_i|$ 维向量：

$$\psi_{i,u}(x_i) = \frac{P_{X_i|U}(x_i|u) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}} \tag{D-26}$$

因此，忽略 ϵ 高阶项后，可将欲求优化问题转化为线性代数问题

$$\underset{\mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]\leqslant 1}{\text{maximize}} \sum_{i=1}^{d} \mathbb{E}_U[\|\boldsymbol{\psi}_{i,U}\|^2] - \mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]. \tag{D-27}$$

为求解 (D-27)，首先注意到 P_{X_i} 及 $P_{X_i|U}$ 分别为 P_{X^d} 及 $P_{X^d|U}$ 对应的边缘分布，因此 $\boldsymbol{\phi}_u$ 与 $\boldsymbol{\psi}_{i,u}$ 满足

$$\psi_{i,u}(x_i) = \sum_{x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_d} \frac{\sqrt{P_{X^d}(x^d)}}{\sqrt{P_{X_i}(x_i)}} \phi_u(x^d).$$

将上式表示为矩阵形式，得 $\boldsymbol{\psi}_{i,u} = \mathbf{B}_i \cdot \boldsymbol{\phi}_u$，其中 $\mathbf{B}_i$ 定义由(D-18) 给出。

因此，根据

$$\sum_{i=1}^{d} \mathbb{E}_U[\|\boldsymbol{\psi}_{i,U}\|^2] = \sum_{i=1}^{d} \mathbb{E}_U[\|\mathbf{B}_i \cdot \boldsymbol{\phi}_U\|^2] = \mathbb{E}_U[\|\mathbf{B}_0 \cdot \boldsymbol{\phi}_U\|^2],$$

可将 (D-27) 表示为

$$\underset{\mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2]\leqslant 1}{\text{maximize}} \ \mathbb{E}_U[\|\mathbf{B}_0 \cdot \boldsymbol{\phi}_U\|^2] - \mathbb{E}_U[\|\boldsymbol{\phi}_U\|^2], \tag{D-28}$$

其中 $\mathbf{B}_0$ 定义由 (D-19) 给出。此外，由于 $\boldsymbol{\phi}_u$ 为信息向量，由定义可知

$$\sum_{x^d} \sqrt{P_{X^d}(x^d)} \phi_u(x^d) = 0, \tag{D-29}$$

从而 $\boldsymbol{\phi}_U$ 正交于由诸 $\sqrt{P_{X^d}(x^d)}$ 构成的 $\left(|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdots |\mathcal{X}_d|\right)$ 维向量 $\boldsymbol{\phi}^{(0)}$。特别地，根据 [35] 可知 $\boldsymbol{\phi}^{(0)}$ 为矩阵 $\mathbf{B}_0$ 对应于最大奇异值 $\sigma_0 = \sqrt{d}$ 的右奇异向量，且相应的左奇异向量为 $\boldsymbol{\psi}^{(0)}$。由 (D-20) 知 $\mathbf{B}_0$ 的第二大奇异值为 $\sigma_1 = \sqrt{\lambda^{(1)}} \geqslant 1$，且 (D-28) 的最优解满足对任意 u，$\boldsymbol{\phi}_{U=u}$ 与 $\mathbf{B}_0$ 第二大奇异值对应的右奇异向量平行。

注意到 $\mathbf{B}_0$ 次大奇异值对应的左奇异向量的计算要比其右奇异向量计算简单，且该左奇异向量也是矩阵 $\mathbf{B}_0\mathbf{B}_0^{\mathsf{T}} = \mathbf{B}$ 的次大奇异值所对应的左奇异向量，即 $\boldsymbol{\psi}^{(1)}$。从而，$\mathbf{B}_0$ 的第二大奇异值对应的右奇异向量 $\boldsymbol{\phi}^{(1)}$ 为

$$\begin{aligned}
\phi^{(1)}(x^d) &= \frac{1}{\sqrt{\lambda^{(1)}}} (\mathbf{B}_0^{\mathsf{T}} \boldsymbol{\psi}^{(1)})(x^d) \\
&= \frac{1}{\sqrt{\lambda^{(1)}}} \cdot \left(\sqrt{P_{X^d}(x^d)} \sum_{i=1}^{d} \frac{\psi_i^{(1)}(x_i)}{\sqrt{P_{X_i}(x_i)}} \right) \\
&= \sqrt{P_{X^d}(x^d)} \cdot \left(\frac{1}{\sqrt{\lambda^{(1)}}} \sum_{i=1}^{d} f_i^{(1)}(x_i) \right),
\end{aligned} \tag{D-30}$$

其中 $(\mathbf{B}_0^{\mathsf{T}} \boldsymbol{\psi}^{(1)})(x^d)$ 为向量 $\mathbf{B}_0^{\mathsf{T}} \boldsymbol{\psi}^{(1)}$ 的第 x^d 个元素。由于所有的 $\boldsymbol{\phi}_{U=u}$ 均与 $\boldsymbol{\phi}^{(1)}$ 平行，

存在函数 $h: U \mapsto \mathbb{R}$ 使得

$$P_{X^d|U}(x^d|u) = P_{X^d}(x^d)\left(1 + \frac{\epsilon h(u)}{\sqrt{\lambda^{(1)}}}\sum_{i=1}^{d} f_i^{(1)}(x_i)\right) + o(\epsilon),$$

其中 $o(\epsilon)$ 项来自 (D-24) 中的局部近似。因此，所考察优化问题对应的最优联合分布可表示为

$$P_{UX^d}(u, x^d) = P_U(u)P_{X^d}(x^d)\left(1 + \frac{\epsilon h(u)}{\sqrt{\lambda^{(1)}}}\sum_{i=1}^{d} f_i^{(1)}(x_i)\right) + o(\epsilon). \tag{D-31}$$

将(D-31)两边对所有的 $u \in U$ 求和，可得 $\sum_{u\in U} P_U(u)h(u) = 0$，从而可知 $h(U)$ 为零均值函数。此外，由 (D-24) 可求得方差 $\mathbb{E}[h^2(U)] = 1$。最后，注意到当 δ 很小时，指数族 $\mathcal{P}_{\text{exp}}^{(\delta)}$ 可表示为

$$\mathcal{P}_{\text{exp}}^{(\delta)} = \left\{ P_U(u)P_{X^d}(x^d)\cdot\left(1 + \frac{\sqrt{2\delta}h(u)}{\sqrt{\lambda^{(1)}}}\sum_{i=1}^{d} f_i^{(1)}(x_i)\right) + o\left(\sqrt{\delta}\right) : h \in \mathcal{H}_\delta \right\}.$$

对比 (D.8) 以及 (D-31)，并结合 $\delta = \frac{1}{2}\epsilon^2$ 可得欲证结论。

D.9　定理 6.3 的证明

首先给出如下引理 (可见 [63] p. 248 中引理 4.3.39)。

引理 D.2： 对任意给定的 $k_1 \times k_2$ 矩阵 $\mathbf{A}$ 及 $k \in \{1, \cdots, \min\{k_1, k_2\}\}$，有

$$\max_{\mathbf{M}\in\mathbb{R}^{k_2\times k}} \|\mathbf{AM}\|_{\text{F}}^2 = \sum_{i=1}^{k} \sigma_i^2, \tag{D-32}$$

其中 $\|\cdot\|_{\text{F}}$ 表示 Frobenius 范数，且 $\sigma_1 \geqslant \cdots \geqslant \sigma_{\min\{m,n\}}$ 为 $\mathbf{A}$ 的奇异值。此外，当 $\mathbf{M} = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_k\end{bmatrix}\mathbf{Q}$ 时，(D-32) 可取得最大值，其中对任意 $i = 1, \cdots, \min\{m, n\}$，$\mathbf{v}_i$ 为 $\mathbf{A}$ 与 σ_i 对应的右奇异向量，且 $\mathbf{Q} \in \mathbb{R}^{k\times k}$ 为任意的正交矩阵。

与定理 6.2 类似，我们首先将 δ 替代为 $\frac{1}{2}\epsilon^2$ 并且定义条件分布 $P_{X^d|U^k=u^k}$ 所对应的 $(|\mathcal{X}_1|\cdot|\mathcal{X}_2|\cdots|\mathcal{X}_d|)$ 维信息向量 $\boldsymbol{\phi}_{u^k}$ 使其满足

$$P_{X^d|U^k}(x^d|u^k) = P_{X^d}(x^d) + \epsilon\sqrt{P_{X^d}(x^d)}\boldsymbol{\phi}_{u^k}(x^d), \tag{D-33}$$

则根据 K-L 散度的二阶 Taylor 展开有

$$I(U^k; X^d) = \mathbb{E}_{U^k}[D(P_{X^d|U^k}\|P_{X^d})]$$

$$= \frac{1}{2}\epsilon^2 \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\phi}_{U^k}\right\|^2\right] + o(\epsilon^2).$$

类似地，对 $i = 1, \cdots, k$，条件分布 $P_{X^d|U_i=u_i}$ 可表为

$$P_{X^d|U_i}(x^d|u_i) = P_{X^d}(x^d) + \epsilon\sqrt{P_{X^d}(x^d)}\boldsymbol{\phi}_{u_i}(x^d), \tag{D-34}$$

从而

$$I(U_i; X^d) = \frac{1}{2}\epsilon^2 \mathbb{E}_{U_i}\left[\left\|\boldsymbol{\phi}_{U_i}\right\|^2\right] + o(\epsilon^2).$$

忽略 ϵ 的高阶项，可将第一个约束条件化简为

$$1 \geqslant \mathbb{E}_{U_1}\left[\left\|\boldsymbol{\phi}_{U_i}\right\|^2\right] \geqslant \cdots \geqslant \mathbb{E}_{U_k}\left[\left\|\boldsymbol{\phi}_{U_i}\right\|^2\right].$$

此外，根据 $\boldsymbol{U}^k$ 的独立性及条件独立性，$\boldsymbol{\phi}_{u^k}$ 与 $\boldsymbol{\phi}_{u_i}$ 满足

$$\boldsymbol{\phi}_{u^k} = \sum_{i=1}^{k} \boldsymbol{\phi}_{u_i} + o(1) \tag{D-35}$$

以及

$$\langle \boldsymbol{\phi}_{u_i}, \boldsymbol{\phi}_{u_j} \rangle = 0, \quad \text{对任意 } i \neq j, u_i \in \mathcal{U}_i, u_j \in \mathcal{U}_j. \tag{D-36}$$

实际上，由

$$\begin{aligned} P_{X_d|U^k}(x^d|u^k) &= \frac{P_{X^d}(x^d)P_{U^k|X^d}(u^k|x^d)}{P_{U^k}(u^k)} \\ &= P_{X^d}(x^d)\prod_{i=1}^{k}\frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)}, \end{aligned}$$

可得

$$\frac{P_{X_d|U^k}(x^d|u^k)}{P_{X^d}(x^d)} = \prod_{i=1}^{k}\frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)} = \prod_{i=1}^{k}\frac{P_{X^d|U_i}(x^d|u_i)}{P_{X^d}(x^d)}. \tag{D-37}$$

将 (D-33) 及 (D-34) 代入 (D-37) 可推出

$$1 + \epsilon\frac{\boldsymbol{\phi}_{u^k}(x_1^d)}{\sqrt{P_{X^d}(x^d)}} = \prod_{i=1}^{k}\left[1 + \epsilon\frac{\boldsymbol{\phi}_{u_i}(x_1^d)}{\sqrt{P_{X^d}(x^d)}}\right].$$

比较等式两边 ϵ 项，得 (D-35)。

为证明 (D-36), 注意到根据 (D-34)，对任意 $i \neq j$，及 $u_i \in \mathcal{U}_i$，$u_j \in \mathcal{U}_j$，有

$$\epsilon^2 \left\langle \boldsymbol{\phi}_{u_i}, \boldsymbol{\phi}_{u_j} \right\rangle = \epsilon^2 \sum_{x^d} \phi_{u_i}(x^d)\phi_{u_j}(x^d) \tag{D-38}$$

$$= \sum_{x_1^d} \Big(\frac{1}{P_{X^d}(x^d)} \cdot [P_{X^d|U_i}(x^d|u_i) - P_{X^d}(x^d)] \cdot [P_{X^d|U_j}(x^d|u_i) - P_{X^d}(x^d)] \Big) \tag{D-39}$$

$$= \sum_{x_1^d} \frac{P_{X^d|U_i}(x^d|u_i) P_{X^d|U_j}(x^d|u_j)}{P_{X^d}(x^d)} - 1 \tag{D-40}$$

$$= \sum_{x_1^d} P_{X^d}(x^d) \cdot \frac{P_{U_i|X^d}(u_i|x^d)}{P_{U_i}(u_i)} \cdot \frac{P_{U_j|X^d}(u_j|x^d)}{P_{U_j}(u_j)} - 1$$

$$= \frac{1}{P_{U_iU_j}(u_i,u_j)} \sum_{x_1^d} P_{X^d}(x^d) P_{U_iU_j|X^d}(u_i,u_j|x^d) - 1 \tag{D-41}$$

$$= 0,$$

其中 (D-41) 利用了 U_i 与 U_j 的独立性及条件独立性。

此外，目标函数 $\mathcal{L}(X^d|U^k)$ 可表示为 [参见 (D-25)]

$$\begin{aligned} & D(P_{X^d} \| P_{X_1} \cdots P_{X_d}) - D(P_{X^d} \| P_{X_1} \cdots P_{X_d}|U^k) \\ &= \sum_{i=1}^{d} I(U^k; X_i) - I(U^k; X^d) \\ &= \sum_{i=1}^{d} I(U^k; X_i) - \sum_{j=1}^{k} I(U_j; X^d), \end{aligned} \tag{D-42}$$

其中最后的等式依据

$$I(U^k; X^d) = \mathbb{E}_{U^kX^d} \left[\log \frac{P_{U^k|X^d}(U^k|X^d)}{P_{U^k}(U^k)} \right] \tag{D-43}$$

$$= \mathbb{E}_{U^kX^d} \left[\sum_{j=1}^{k} \log \frac{P_{U_j|X^d}(U_j|X^d)}{P_{U_j}(U_j)} \right] \tag{D-44}$$

$$= \sum_{j=1}^{k} I(U_j; X^d), \tag{D-45}$$

且 (D-44) 基于 $U_1, \cdots, U_k$ 的独立性及关于 X^d 的条件独立性。

对所有 i，互信息 $I(\boldsymbol{U}^k; X_i)$ 可近似为

$$I(\boldsymbol{U}^k; X_i) = \frac{1}{2}\epsilon^2 \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\psi}_{i,U^k}\right\|^2\right] + o(\epsilon^2),$$

其中对任意 $\boldsymbol{U}^k = u^k$，$\boldsymbol{\psi}_{i,u^k}$ 定义为 $|\mathcal{X}_i|$ 维信息向量：

$$\psi_{i,u^k}(x_i) = \frac{P_{X_i|U^k}(x_i|u^k) - P_{X_i}(x_i)}{\epsilon\sqrt{P_{X_i}(x_i)}}.$$

因此当忽略 ϵ 高阶项时，全相关优化问题可化简为

$$\underset{\boldsymbol{\phi}_{u^k}}{\text{maximize}} \quad \sum_{i=1}^{d} \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\psi}_{i,U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] \tag{D-46a}$$

$$\text{subject to} \quad 1 \geqslant \mathbb{E}_{U_1}\left[\left\|\boldsymbol{\phi}_{U_1}\right\|^2\right] \geqslant \cdots \geqslant \mathbb{E}_{U_k}\left[\left\|\boldsymbol{\phi}_{U_k}\right\|^2\right] \tag{D-46b}$$

$$\langle \boldsymbol{\phi}_{u_i}, \boldsymbol{\phi}_{u_j} \rangle = 0, \quad i \neq j, u_i \in \mathcal{U}_i, u_j \in \mathcal{U}_j \tag{D-46c}$$

$$\left\langle \boldsymbol{\phi}_{u_j}, \boldsymbol{\phi}^{(0)} \right\rangle = 0, \forall\, u_j \in \mathcal{U}_j, j = 1, \cdots, k \tag{D-46d}$$

$$\boldsymbol{\phi}_{u^k} = \sum_{j=1}^{k} \boldsymbol{\phi}_{u_j}, \forall\, u^k \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_k. \tag{D-46e}$$

为求解 (D-46)，首先注意到 $\boldsymbol{\psi}_{i,U^k} = \mathbf{B}_i \boldsymbol{\phi}_{U^k}$, 其中 $\mathbf{B}_i$ 定义由 (D-18) 给出。目标函数 (D-46a) 可表示为

$$\begin{aligned}
&\sum_{i=1}^{d} \mathbb{E}_{U^k}\left[\left\|\boldsymbol{\psi}_{i,U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{U^k}\left[\left\|\mathbf{B}_i \boldsymbol{\phi}_{U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(D-47)} \\
&= \mathbb{E}_{U^k}\left[\left\|\mathbf{B}_0 \boldsymbol{\phi}_{U^k}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(D-48)} \\
&= \mathbb{E}_{U^k}\left[\left\|\sum_{j=1}^{k} \mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(D-49)} \\
&= \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right\|^2\right] - \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(D-50)} \\
&= \sum_{j=1}^{k} \mathbb{E}_{U_j}\left[\left\|\mathbf{B}_0 \boldsymbol{\phi}_{U_j}\right\|^2 - \left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] && \text{(D-51)}
\end{aligned}$$

其中 $\mathbf{B}_0$ 定义参见 (D-19)。为导出 (D-50)，只需注意到对 $i \neq j$，有

$$\mathbb{E}_{U^k}\left[\boldsymbol{\phi}_{U_i}^{\mathsf{T}}\mathbf{B}_0^{\mathsf{T}}\mathbf{B}_0\boldsymbol{\phi}_{U_j}\right] = \left(\mathbb{E}_{U_i}\left[\boldsymbol{\phi}_{U_i}\right]\right)^{\mathsf{T}}\mathbf{B}_0^{\mathsf{T}}\mathbf{B}_0\left(\mathbb{E}_{U_j}\left[\boldsymbol{\phi}_{U_j}\right]\right) = 0,$$

其中第一个等号基于 U_i 与 U_j 相互独立，第二个等号基于 $\mathbb{E}_{U_i}\left[\boldsymbol{\phi}_{U_i}\right] = 0$.

为最大化 (D-51), 诸 $\boldsymbol{\phi}_{u_i}, u_i \in \mathcal{U}_i$ 应平行，否则可在保持 $\mathbb{E}_{U_i}\left[\|\boldsymbol{\phi}_{U_i}\|^2\right]$ 不变的前提下令所有 $\boldsymbol{\phi}_{u_i}$ 都平行于

$$\underset{\boldsymbol{\phi}_{u_i}\,:\,u_i\in\mathcal{U}_i}{\arg\max}\ \frac{\|\mathbf{B}_0\boldsymbol{\phi}_{u_i}\|^2}{\|\boldsymbol{\phi}_{u_i}\|^2},$$

从而目标函数值将增大，与最优性矛盾。

因此，对任意 i 以及 $u_i \in \mathcal{U}_i$，$\boldsymbol{\phi}_{u_i}$ 可表示为

$$\boldsymbol{\phi}_{u_i} = h_i(u_i)\boldsymbol{\phi}_i, \tag{D-52}$$

其中 $h_i : \mathcal{U}_i \mapsto \mathbb{R}$ 且 $\boldsymbol{\phi}_i$ 为单位向量。从而可得 $\mathbb{E}_{U_i}[\boldsymbol{\phi}_{U_i}] = \mathbb{E}_{U_i}[h_i(U_i)]\boldsymbol{\phi}_i = 0$ 及

$$\mathbb{E}_{U_i}[\|\boldsymbol{\phi}_{U_i}\|^2] = \mathbb{E}_{U_i}[h_i^2(U_i)], \tag{D-53a}$$

$$\mathbb{E}_{U_i}[\|\mathbf{B}_0\boldsymbol{\phi}_{U_i}\|^2] = \mathbb{E}_{U_i}[h_i^2(U_i)]\|\mathbf{B}_0\boldsymbol{\phi}_i\|^2. \tag{D-53b}$$

于是约束条件 (D-46b) 可化简为

$$1 \geqslant \mathbb{E}_{U_1}[h_1^2(U_1)] \geqslant \cdots \geqslant \mathbb{E}_{U_k}[h_k^2(U_k)]. \tag{D-54}$$

此外，由 (D-53) 可知 $\mathbb{E}_{U_i}[\|\mathbf{B}_0\boldsymbol{\phi}_{U_i}\|^2] - \mathbb{E}_{U_i}[\|\boldsymbol{\phi}_{U_i}\|^2] = \mathbb{E}_{U_i}[h_i^2(U_i)]\left[\|\mathbf{B}_0\boldsymbol{\phi}_i\|^2 - 1\right]$, 故最大化 (D-51) 的 h_i 应满足

$$\mathbb{E}_{U_i}[h_i^2(U_i)] = \begin{cases} 1 & \text{若 } \|\mathbf{B}_0\boldsymbol{\phi}_i\|^2 > 1 \\ 0 & \text{其他情况.} \end{cases}$$

由 (D-54) 知存在 $k_0 \in \{1, \cdots, k\}$ 使得

$$\mathbb{E}_{U_i}\left[h_i^2(U_i)\right] = \begin{cases} 1 & i = 1, \cdots, k_0 \\ 0 & i > k_0, \end{cases} \tag{D-55}$$

于是目标函数 (D-51) 可化简为

$$\sum_{j=1}^{k}\mathbb{E}_{U_j}\left[\left\|\mathbf{B}_0\boldsymbol{\phi}_{U_j}\right\|^2 - \left\|\boldsymbol{\phi}_{U_j}\right\|^2\right] = \sum_{j=1}^{k_0}\|\mathbf{B}_0\boldsymbol{\phi}_i\|^2 - k_0$$

$$= \left\|\mathbf{B}_0\boldsymbol{\Phi}_0\right\|_{\mathrm{F}}^2 - k_0,$$

其中 $\boldsymbol{\Phi}_0 \triangleq \begin{bmatrix}\boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_{k_0}\end{bmatrix}$.

因此优化问题 (D-46) 等价于

$$\underset{\boldsymbol{\Phi}_0}{\text{maximize}} \ \left\|\mathbf{B}_0\boldsymbol{\Phi}_0\right\|_{\mathrm{F}}^2 - k_0 \tag{D-56a}$$

$$\text{subject to } \boldsymbol{\Phi}_0^{\mathsf{T}}\boldsymbol{\Phi}_0 = \mathbf{I}_{k_0} \tag{D-56b}$$

$$\boldsymbol{\Phi}_0^{\mathsf{T}}\boldsymbol{\phi}^{(0)} = \mathbf{0}_{k_0}, \tag{D-56c}$$

其中 $\mathbf{I}_{k_0}$ 为 k_0 阶单位阵，$\mathbf{0}_{k_0}$ 为 $\mathbb{R}^{k_0}$ 中的零向量。又由于 $\boldsymbol{\phi}^{(0)}$ 为 $\mathbf{B}_0$ 最大奇异值对应的右奇异向量，(D-56) 可进一步化简为

$$\underset{\boldsymbol{\Phi}_0}{\text{maximize}} \ \left\|\tilde{\mathbf{B}}_0\boldsymbol{\Phi}_0\right\|_{\mathrm{F}}^2 - k_0 \tag{D-57a}$$

$$\text{subject to } \boldsymbol{\Phi}_0^{\mathsf{T}}\boldsymbol{\Phi}_0 = \mathbf{I}_{k_0}, \tag{D-57b}$$

其中 $\tilde{\mathbf{B}}_0 \triangleq \mathbf{B}_0 - \sqrt{\lambda^{(0)}}\boldsymbol{\psi}^0\left(\boldsymbol{\phi}^{(0)}\right)^{\mathsf{T}}$.

由引理 D.2 可知，(D-57) 的最优值为

$$\sum_{i=1}^{k_0}\lambda^{(i)} - k_0 = \sum_{i=1}^{k_0}\left[\lambda^{(i)} - 1\right]. \tag{D-58}$$

为最大化 (D-58)，k_0 应取为使得 $\lambda^{(i)} > 1$ 的最大的 i，亦即 $k_0 = \min\{k, k^*\}$。此外，最优的 $\boldsymbol{\Phi}_0$ 为 $\boldsymbol{\Phi}_0 = \begin{bmatrix}\boldsymbol{\phi}^{(1)} & \cdots & \boldsymbol{\phi}^{(k_0)}\end{bmatrix}\mathbf{Q}$，其中 $\mathbf{Q} \in \mathbb{R}^{k_0\times k_0}$ 满足 $\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}_{k_0}$。因此

$$\boldsymbol{\phi}_\ell = \sum_{j=1}^{k_0} q_{j\ell}\boldsymbol{\phi}^{(j)}.$$

与 (D-30) 同理，$\boldsymbol{\phi}^{(j)}$ 可表示为

$$\frac{\phi^{(j)}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \frac{1}{\sqrt{\lambda^{(j)}}}\sum_{i=1}^{d} f_i^{(j)}(x_i),$$

故

$$\frac{\phi_\ell(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{j=1}^{k_0} q_{j\ell}\cdot\frac{\phi^{(j)}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{j=1}^{k_0}\frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}}\sum_{i=1}^{d} f_i^{(j)}(x_i).$$

从而由 (D-52) 可得对任意 $\ell = 1, \cdots, k_0$，有

$$\frac{\phi_{u_\ell}(x^d)}{\sqrt{P_{X^d}(x^d)}} = h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i). \tag{D-59}$$

又由 (D-46e) 可知

$$\boldsymbol{\phi}_{u^k} = \sum_{\ell=1}^{k} \boldsymbol{\phi}_{u_\ell} = \sum_{\ell=1}^{k_0} \boldsymbol{\phi}_{u_\ell},$$

其中第二个等号基于 $\boldsymbol{\phi}_{u_\ell} = \mathbf{0}$ $(\ell > k_0)$，后者可利用 (D-52) 及 (D-55) 得出。

因此有

$$\frac{\phi_{u^k}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{\ell=1}^{k_0} \frac{\phi_{u_\ell}(x^d)}{\sqrt{P_{X^d}(x^d)}} = \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i),$$

从而可得

$$\begin{aligned} P_{X^d|U^k}(x^d|u^k) &= P_{X^d}(x^d) \left[1 + \epsilon \frac{\phi_{u^k}(x^d)}{\sqrt{P_{X^d}(x^d)}} \right] + o(\epsilon) \\ &= P_{X^d}(x^d) \left[1 + \epsilon \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] + o(\epsilon) \end{aligned}$$

以及

$$\begin{aligned} &P_{X^d U^k}(x^d, u^k) \\ &\quad = P_{X^d}(x^d) \left[\prod_{j=1}^{k} P_{U_j}(u_j) \right] \cdot \left[1 + \epsilon \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] + o(\epsilon). \end{aligned} \tag{D-60}$$

最后，注意到当 δ 很小时，指数族 $\mathcal{P}_{\exp,k}^{(\delta)}$ 可表示为

$$\begin{aligned} \mathcal{P}_{\exp,k}^{(\delta)} = \Bigg\{ & P_{X^d}(x^d) \left[\prod_{j=1}^{k} P_{U_j}(u_j) \right] \cdot \left[1 + \sqrt{2\delta} \sum_{\ell=1}^{k_0} h_\ell(u_\ell) \sum_{j=1}^{k_0} \frac{q_{j\ell}}{\sqrt{\lambda^{(j)}}} \sum_{i=1}^{d} f_i^{(j)}(x_i) \right] \\ & : \ h_\ell \in \mathcal{H}_\ell, \mathbf{Q} = [q_{ij}]_{k_0 \times k_0}, \mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}_{k_0} \Bigg\}. \end{aligned} \tag{D-61}$$

对比 (D-61) 及 (D-60)，并根据 $\delta = \frac{1}{2}\epsilon^2$ 即得欲证结论。

D.10　联合相关最大化

给定函数 $\underline{f}_i : \mathcal{X}_i \mapsto \mathbb{R}^k, i = 1, \cdots, d$, 定义 $\boldsymbol{\Psi}_i \in \mathbb{R}^{|\mathcal{X}_i| \times k}$ 使得 $\boldsymbol{\Psi}_i$ 的行向量分别为 $\sqrt{P_{X_i}(x_i)}\underline{f}_i^{\mathsf{T}}(x_i)$，$x_i \in \mathcal{X}_i$。在此基础上，定义 $m \times k$ 矩阵 $\boldsymbol{\Psi}$ 为 $\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_1^{\mathsf{T}} & \cdots & \boldsymbol{\Psi}_d \end{bmatrix}$, 则优化问题 (6-27) 可表示成

$$\underset{\boldsymbol{\Psi}:\, \boldsymbol{\Psi} \in \mathbb{R}^{m \times k}}{\text{maximize}} \quad \operatorname{tr}\left\{ \boldsymbol{\Psi}^{\mathsf{T}} \mathbf{B} \boldsymbol{\Psi} \right\} \tag{D-62a}$$

$$\text{subject to} \quad \boldsymbol{\Psi}_i^{\mathsf{T}} \mathbf{v}_i = \mathbf{0}_k, \ \forall\, i \tag{D-62b}$$

$$\boldsymbol{\Psi}^{\mathsf{T}} \boldsymbol{\Psi} = \mathbf{I}_k, \tag{D-62c}$$

其中 $\mathbf{0}_k$ 为 $\mathbb{R}^k$ 中的零向量，且 $\mathbf{I}_k$ 为 $k \times k$ 单位阵。为给出 (6-27) 与 (D-62) 的等价性，注意到

$$\begin{aligned} \boldsymbol{\Psi}^{\mathsf{T}} \boldsymbol{\Psi} = \sum_{i=1}^{d} \boldsymbol{\Psi}_i^{\mathsf{T}} \boldsymbol{\Psi}_i &= \sum_{i=1}^{d} \sum_{x_i \in \mathcal{X}_i} P_{X_i}(x_i) \underline{f}_i(x_i) \underline{f}_i^{\mathsf{T}}(x_i) \\ &= \sum_{i=1}^{d} \mathbb{E}\left[\underline{f}_i(X_i) \underline{f}_i^{\mathsf{T}}(X_i) \right] \\ &= \mathbb{E}\left[\sum_{i=1}^{d} \underline{f}_i(X_i) \underline{f}_i^{\mathsf{T}}(X_i) \right] \end{aligned}$$

及

$$\begin{aligned} \operatorname{tr}\left\{ \boldsymbol{\Psi}^{\mathsf{T}} \mathbf{B} \boldsymbol{\Psi} \right\} &= \sum_{i=1}^{d} \sum_{j=1}^{d} \operatorname{tr}\left\{ \boldsymbol{\Psi}_i^{\mathsf{T}} \mathbf{B}_{ij} \boldsymbol{\Psi}_j \right\} \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} \operatorname{tr}\left\{ \mathbb{E}\left[\underline{f}_i(X_i) \underline{f}_j^{\mathsf{T}}(X_j) \right] \right\} \\ &= \sum_{i=1}^{d} \operatorname{tr}\left\{ \mathbb{E}\left[\underline{f}_i(X_i) \underline{f}_i^{\mathsf{T}}(X_i) \right] \right\} + \sum_{i \neq j} \operatorname{tr}\left\{ \mathbb{E}\left[\underline{f}_i(X_i) \underline{f}_j^{\mathsf{T}}(X_j) \right] \right\} \\ &= \operatorname{tr}\left\{ \sum_{i=1}^{d} \mathbb{E}\left[\underline{f}_i(X_i) \underline{f}_i^{\mathsf{T}}(X_i) \right] \right\} + \sum_{i \neq j} \operatorname{tr}\left\{ \mathbb{E}\left[\underline{f}_i(X_i) \underline{f}_j^{\mathsf{T}}(X_j) \right] \right\} \\ &= k + \mathbb{E}\left[\sum_{i \neq j} \underline{f}_i^{\mathsf{T}}(X_i) \underline{f}_j(X_j) \right]. \end{aligned}$$

由引理 6.1 知对所有的 $k < m - d$，(D-62) 的解可表示为 $\boldsymbol{\Psi}^* = \begin{bmatrix} \boldsymbol{\psi}^{(1)} & \cdots & \boldsymbol{\psi}^{(k)} \end{bmatrix} \mathbf{Q}$, 其中 $\mathbf{Q} \in \mathbb{R}^{k \times k}$ 为正交阵。故 (6-27) 的最优解对应于 $f_i^{(\ell)}$，其中 $i = 1, \cdots, d$，$\ell = 1, \cdots, k$。

D.11　比特公共模式提取

首先定义 $\ell_{\max}$ 为使 $w(\mathcal{J}_\ell) > 0$ 的最大的 ℓ，亦即 $\ell_{\max} \triangleq \max\{\ell : 0 \leqslant \ell \leqslant 2^r - 1, w(\mathcal{J}_\ell) > 0\}$，则 $w(\mathcal{J}_\ell) > 0$ 等价于 $\ell \leqslant \ell_{\max}$，且 (6-29) 可等价地表示为

$$\lambda^{(\ell)} = w(\mathcal{J}_\ell), \quad \ell \leqslant \ell_{\max}, \tag{D-63}$$

以及

$$\lambda^{(\ell)} = 0, \quad \ell > \ell_{\max}. \tag{D-64}$$

注意到 (6-22) 建立了诸函数 $f_i^{(\ell)}$ $(i = 1, \cdots, d)$ 与向量 $\boldsymbol{\psi}^{(\ell)}$ 的一一对应关系。基于该关系，可使用 $\tilde{\boldsymbol{\psi}}^{(\ell)}$ 表示由 (6-30) 定义的 $f_i^{(\ell)}$。

接下来的证明可分为如下两步：首先，我们验证 $\tilde{\boldsymbol{\psi}}^{(\ell)}$ $(\ell = 0, \cdots, \ell_{\max})$ 为 $\mathbf{B}$ 分别关于特征值 $w(\mathcal{J}_\ell)$ $(\ell = 0, \cdots, \ell_{\max})$ 的 $(\ell_{\max} + 1)$ 个正交的特征向量，即对任意 $0 \leqslant \ell \leqslant \ell_{\max}$ 与 $0 \leqslant \ell' \leqslant \ell_{\max}$，诸 $\tilde{\boldsymbol{\psi}}^{(\ell)}$ 满足

$$\mathbf{B}\tilde{\boldsymbol{\psi}}^{(\ell)} = w(\mathcal{J}_\ell)\tilde{\boldsymbol{\psi}}^{(\ell)} \quad \text{及} \quad \left\langle \tilde{\boldsymbol{\psi}}^{(\ell)}, \tilde{\boldsymbol{\psi}}^{(\ell')} \right\rangle = \delta_{\ell\ell'}, \tag{D-65}$$

其中 $\delta_{\ell\ell'}$ 为 Kronecker delta 记号。在此基础上，只需验证 $\mathbf{B}$ 的其它所有特征值均为 0 [参见(D-64)]。

为此，首先将 (D-65) 表示成 $f_i^{(\ell)}$ 的形式，得

$$\sum_{j=1}^{d} \mathbb{E}\left[f_j^{(\ell)}(X_j) \middle| X_i \right] = w(\mathcal{J}_\ell) f_i^{(\ell)}(X_i), \quad 1 \leqslant i \leqslant d, \tag{D-66}$$

以及

$$\sum_{i=1}^{d} \mathbb{E}\left[f_i^{(\ell)}(X_i) f_i^{(\ell')}(X_i) \right] = \delta_{\ell\ell'}. \tag{D-67}$$

根据 [参见 (6-28)]

$$\sum_{i=1}^{d} \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_i\}} = \sum_{j=1}^{d} \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_j\}} = w(\mathcal{J}_\ell),$$

只需验证

$$\mathbb{E}\left[f_j^{(\ell)}(X_j) \middle| X_i \right] = f_i^{(\ell)}(X_i) \cdot \mathbb{1}_{\{\mathcal{J}_\ell \subset \mathcal{I}_j\}}, \quad 1 \leqslant i, j \leqslant d, \tag{D-68}$$

以及

$$\mathbb{E}\left[f_i^{(\ell)}(X_i)f_i^{(\ell')}(X_i)\right]=\frac{\mathbb{1}_{\{\mathcal{J}_\ell\subset\mathcal{I}_i\}}}{w(\mathcal{J}_\ell)}\cdot\delta_{\ell\ell'},\quad 1\leqslant i\leqslant d. \tag{D-69}$$

为导出 (D-68)，注意到若 $\mathcal{J}_\ell\not\subset\mathcal{I}_j$，由 (6-30) 可得 $f_j(X_j)=0$，从而 (D-68) 成立；否则可得到 $\mathcal{J}_\ell\subset\mathcal{I}_j$ 及

$$\mathbb{E}\left[f_j^{(\ell)}(X_j)\middle|X_i\right]=\frac{1}{\sqrt{w(\mathcal{J}_\ell)}}\mathbb{E}\left[\prod_{s\in\mathcal{J}_\ell}b_s\middle|X_i\right]. \tag{D-70}$$

因为 $X_i=b_{\mathcal{I}_i}$ 由下标在 $\mathcal{I}_i$ 中的诸 b_s 构成，故

$$\mathbb{E}\left[\prod_{s\in\mathcal{J}_\ell}b_s\middle|X_i\right]=\begin{cases}\prod\limits_{s\in\mathcal{J}_\ell}b_s & \text{若 }\mathcal{J}_\ell\subset\mathcal{I}_i\\ 0 & \text{其他情况,}\end{cases}$$

由此推出

$$\begin{aligned}\mathbb{E}\left[f_j^{(\ell)}(X_j)\middle|X_i\right]&=\frac{1}{\sqrt{w(\mathcal{J}_\ell)}}\mathbb{E}\left[\prod_{s\in\mathcal{J}_\ell}b_s\middle|X_i\right]\\&=\begin{cases}\dfrac{1}{\sqrt{w(\mathcal{J}_\ell)}}\prod\limits_{s\in\mathcal{J}_\ell}b_s & \text{若 }\mathcal{J}_\ell\subset\mathcal{I}_i\\ 0 & \text{其他情况}\end{cases}\\&=f_i^{(\ell)}(X_i)=f_i^{(\ell)}(X_i)\cdot\mathbb{1}_{\{\mathcal{J}_\ell\subset\mathcal{I}_j\}}.\end{aligned}$$

同理，若 $\ell=\ell'$，(D-69) 可立即由 (6-30) 推出, 从而只需考虑 $\ell\neq\ell'$ 的情形并证明

$$\mathbb{E}\left[f_i^{(\ell)}(X_i)f_i^{(\ell')}(X_i)\right]=0. \tag{D-71}$$

实际上，若有 $\mathcal{J}_\ell\not\subset\mathcal{I}_i$ 或 $\mathcal{J}_{\ell'}\not\subset\mathcal{I}_i$，(D-71) 显然成立；否则有 $\mathcal{J}_\ell\subset\mathcal{I}_i$ 以及 $\mathcal{J}_{\ell'}\subset\mathcal{I}_i$，从而根据 (6-30) 得

$$f_i^{(\ell)}(X_i)f_i^{(\ell')}(X_i)=\frac{1}{\sqrt{w(\mathcal{J}_\ell)w(\mathcal{J}_{\ell'})}}\prod_{j\in\mathcal{J}_\ell\triangle\mathcal{J}_{\ell'}}b_j,$$

其中 “$\triangle$” 表示两个集合的对称差，亦即 $A\triangle B=(A\setminus B)\cup(B\setminus A)$。由此得出

$$\mathbb{E}\left[f_i^{(\ell)}(X_i)f_i^{(\ell')}(X_i)\right]=\frac{1}{\sqrt{w(\mathcal{J}_\ell)w(\mathcal{J}_{\ell'})}}\prod_{j\in\mathcal{J}_\ell\triangle\mathcal{J}_{\ell'}}\mathbb{E}\left[b_j\right]=0.$$

注意到由于 $\mathcal{J}_\ell \neq \mathcal{J}_{\ell'}$，上式中的集合 $(\mathcal{J}_\ell \triangle \mathcal{J}_{\ell'})$ 非空。

最后，为证明 (D-64)，即除 $w(\mathcal{J}_\ell)$ $(\ell = 0, \cdots, \ell_{\max})$ 外所有特征值均为零，首先注意到

$$\begin{aligned}\sum_{\ell=0}^{\ell_{\max}} w(\mathcal{J}_\ell) = \sum_{\ell=0}^{2^r-1} w(\mathcal{J}_\ell) &= \sum_{\mathcal{I}\subset[r]} w(\mathcal{I}) \\ &= \sum_{\mathcal{I}\subset[r]} \sum_{i=1}^{d} \mathbb{1}_{\{\mathcal{I}\subset\mathcal{I}_i\}} \\ &= \sum_{i=1}^{d} \sum_{\mathcal{I}\subset[r]} \mathbb{1}_{\{\mathcal{I}\subset\mathcal{I}_i\}} \\ &= \sum_{i=1}^{d} 2^{|\mathcal{I}_i|} = \sum_{i=1}^{d} |\mathcal{X}_i| = m.\end{aligned}$$

其次，所有特征值的和满足

$$\sum_{\ell=0}^{m-1} \lambda^{(\ell)} = \operatorname{tr}\{\mathbf{B}\} = m.$$

由 引理 6.1 可知，$\mathbf{B}$ 所有特征值均为非负，由此可推出 (D-64)。

D.12　命题 6.5 的证明

首先将 (6-21) 中定义的 $\tilde{\mathbf{B}}$ 表示为分块矩阵

$$\tilde{\mathbf{B}} = \begin{bmatrix} \tilde{\mathbf{B}}_{11} & \tilde{\mathbf{B}}_{12} & \cdots & \tilde{\mathbf{B}}_{1d} \\ \tilde{\mathbf{B}}_{21} & \tilde{\mathbf{B}}_{22} & \cdots & \tilde{\mathbf{B}}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{B}}_{d1} & \tilde{\mathbf{B}}_{d2} & \cdots & \tilde{\mathbf{B}}_{dd} \end{bmatrix}, \tag{D-72}$$

其中 $\tilde{\mathbf{B}}_{ij}$ 为 $(|\mathcal{X}_i| \times |\mathcal{X}_j|)$ 的矩阵。于是 $\left\|\tilde{\mathbf{B}} - \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{T}}\right\|_{\mathrm{F}}^2$ 可写作

$$\begin{aligned}\left\|\tilde{\mathbf{B}} - \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{T}}\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^{d}\sum_{j=1}^{d} \left\|\tilde{\mathbf{B}}_{ij} - \boldsymbol{\Psi}_i\boldsymbol{\Psi}_j^{\mathsf{T}}\right\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} \left[\left\|\tilde{\mathbf{B}}_{ij}\right\|_{\mathrm{F}}^2 - 2\operatorname{tr}\left\{\boldsymbol{\Psi}_i^{\mathsf{T}}\tilde{\mathbf{B}}_{ij}\boldsymbol{\Psi}_j\right\} + \left\|\boldsymbol{\Psi}_i\boldsymbol{\Psi}_j\right\|_{\mathrm{F}}^2\right] \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} \left[\left\|\tilde{\mathbf{B}}_{ij}\right\|_{\mathrm{F}}^2 - 2H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right)\right]\end{aligned}$$

$$= \left\|\tilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 - 2H\left(\underline{f}_1(X_1), \cdots, \underline{f}_d(X_d)\right), \tag{D-73}$$

其中第三个等号基于性质

$$\begin{aligned}\operatorname{tr}\left\{\boldsymbol{\Psi}_i^{\mathsf{T}}\tilde{\mathbf{B}}_{ij}\boldsymbol{\Psi}_j\right\} - \frac{1}{2}\left\|\boldsymbol{\Psi}_i\boldsymbol{\Psi}_j\right\|_{\mathrm{F}}^2 &= \mathbb{E}\left[\underline{f}_i^{\mathsf{T}}(X_i)\underline{f}_j(X_j)\right] - \left(\mathbb{E}\left[\underline{f}_i(X_i)\right]\right)^{\mathsf{T}}\mathbb{E}\left[\underline{f}_j(X_j)\right] \\ &\quad - \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}\left[\underline{f}_i(X_i)\underline{f}_i^{\mathsf{T}}(X_i)\right]\mathbb{E}\left[\underline{f}_j(X_i)\underline{f}_j^{\mathsf{T}}(X_j)\right]\right\} \\ &= H\left(\underline{f}_i(X_i), \underline{f}_j(X_j)\right).\end{aligned}$$

个人简历、在学期间发表的学术论文与研究成果

个人简历

1993 年 12 月 21 日出生于江西省上饶县。

2010 年 9 月考入清华大学电子工程系电子信息科学与技术专业，2014 年 7 月本科毕业并获得工学学士学位。

2014 年 9 月免试进入清华大学电子工程系攻读信息与通信工程工学博士学位至今。

发表的学术论文

[1] Xu X, Huang S L. Maximal correlation regression. IEEE Access, 2020, 8:26591-26601. (SCI 收录, 检索号：LC6TQ, 影响因子：3.745)

[2] Xu X, Huang S L. On the optimal tradeoff between computational efficiency and generalizability of Oja's algorithm. IEEE Access, 2020, 8:102616-102628. (SCI 源刊, 影响因子：3.745)

[3] Xu X, Huang S L, Zheng L, et al. The geometric structure of generalized softmax learning // 2018 IEEE Information Theory Workshop (ITW). IEEE, 2018: 1-5. (EI 收录, 检索号: 20190906562730)

[4] Xu X, Huang S L. On the asymptotic sample complexity of HGR maximal correlation functions in semi-supervised learning // 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2019: 879-886. (EI 收录, 检索号: 20200308042352)

[5] Huang S L, Xu X, Zheng L. An information-theoretic approach to unsupervised feature selection for high-dimensional data. IEEE Journal on Selected Areas in Information Theory, 2020. (国际期刊, 已发表)

[6] Huang S L, Xu X. On the robustness of noisy ACE algorithm and multi-layer residual learning // 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019: 2474-2478. (EI 收录, 检索号: 20194207537583)

[7] Huang S L, Xu X. On the sample complexity of HGR maximal correlation functions // 2019 IEEE Information Theory Workshop (ITW). 2019: 1-5. (EI 收录, 检索号: 20200908238845)

[8] Huang S L, Xu X, Zheng L, et al. An information theoretic interpretation to deep

neural networks // 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019: 1984-1988. (EI 收录, 检索号: 20194207537745)

[9] Xu X, Zhang P, Zhang L. Gotcha: a mobile urban sensing system // Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems. ACM, 2014: 316-317. (EI 收录, 检索号: 20144900299369)

[10] Xu X, Chen X, Liu X, et al. Gotcha II: Deployment of a vehicle-based environmental sensing system // Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM. ACM, 2016: 376-377. (EI 收录, 检索号: 20165303197495)

[11] Xu X, Wang W, Huang S L. On the Sample Complexity of Estimating Small Singular Modes // 2020 IEEE International Symposium on Information Theory (ISIT). IEEE, 2020. (已录用. 国际会议.)

[12] Huang S L, Xu X, Zheng L, et al. A Local Characterization for Wyner Common Information // 2020 IEEE International Symposium on Information Theory (ISIT). IEEE, 2020. (已录用. 国际会议.)

[13] Yang T, Xu X, Guo Q, et al. EV charging behaviour analysis and modelling based on mobile crowdsensing data. IET Generation, Transmission & Distribution, 2017, 11(7):1683-1691. (SCI 收录, 检索号: EW9NQ, 影响因子: 2.862)

[14] Chen X, Xu X, Liu X, et al. HAP: Fine-grained dynamic air pollution map reconstruction by hybrid adaptive particle filter // Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM. ACM, 2016: 336-337. (EI 收录, 检索号: 20165303197659)

[15] Liu X, Xu X, Chen X, et al. Individualized calibration of industrial-grade gas sensors in air quality sensing system // Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems. ACM, 2017: 74. (EI 收录, 检索号: 20183105615732)

[16] Ma R, Xu X, Noh H Y, et al. Generative model based fine-grained air pollution inference for mobile sensing systems // Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems. ACM, 2018: 426-427. (EI 收录, 检索号: 20190806530854)

[17] Chen X, Xu X, Liu X, et al. PGA: Physics guided and adaptive approach for mobile fine-grained air pollution estimation // Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, 2018: 1321-1330. (EI 收录, 检索号: 20185106255213)

[18] Ma R, Xu X, Wang Y, et al. Guiding the data learning process with physical model in air pollution inference // 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 4475-4483. (EI 收录, 检索号: 20191106616437)

[19] Liu X, Chen X, Xu X, et al. Delay effect in mobile sensing system for urban air pollution monitoring // Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems. ACM, 2017: 73. (EI 收录, 检索号: 20183105616428)

[20] Li L, Li Y, Xu X, et al. A maximal correlation embedding method for multilabel human context recognition // Proceedings of the 18th International Conference on Information Processing in Sensor Networks. ACM, 2019: 305-306. (EI 收录, 检索号: 20192307012453)

[21] Li L, Li Y, Xu X, et al. Maximal correlation embedding network for multilabel learning with missing labels // 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019: 393-398. (EI 收录, 检索号: 20193407349277)

[22] Ma R, Liu N, Xu X, et al. A deep autoencoder model for pollution map recovery with mobile sensing networks // Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. ACM, 2019: 577-583. (EI 收录, 检索号: 20194107505801)

[23] Wang L, Wu J, Huang S L, et al. An efficient approach to informative feature extraction from multimodal data // Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 5281-5288.