中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

面向复杂场景的AUC优化理论、方法及应用

作者姓名：杨智勇

指导教师：黄庆明 教授

中国科学院大学

学位类别：工学博士

学科专业：网络空间安全

培养单位：中国科学院信息工程研究所

2021 年 6 月

AUC Optimization for Complex Scenarios: Theory, Method and Application

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

in Cyberspace Security

By

Yang Zhiyong

Supervisor: Professor Huang Qingming

Institute of Information Engineering, Chinese Academy of Sciences

June, 2021

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的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，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对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或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

ROC曲线下面积（AUC）是衡量分类器质量的常用指标，可反映给定分类器在不同阈值下的平均性能，对标签分布和代价不敏感。同时AUC指标也可等价表示为正样本得分高于负样本的概率，因此可反映分类器的排序能力。由于上述特性，AUC在长尾数据分类问题中成为标准性能指标之一，可服务于网络空间公共安全、智能医疗、金融分析、推荐及信息检索等重大应用问题。然而，AUC指标本身针对有监督的二分类问题设计，无法适用于半监督学习、多类别分类学习、多任务学习等实际应用中更为常见的复杂场景，导致无法进行相应AUC优化。随着近年来数据体量的飞速增长及人工智能领域应用的日益复杂化、多样化，现有AUC优化算法的上述局限性势必愈发凸显。如何构建面向复杂场景、复杂需求的AUC优化算法框架成为一个亟需解决的关键问题。鉴于此，本文面向半监督、多分类、多任务三种复杂场景，对AUC优化的方法理论及应用展开了系列研究。总体内容概括如下：

第一，提出一种基于boosting算法的半监督AUC优化算法。现有大多数方法局限于通过单一模型进行半监督AUC优化，对如何通过boosting技术融合多个模型则鲜有涉及。考虑上述局限性，主要研究基于模型集成的半监督AUC优化方法。具体而言，提出一种基于boosting算法的半监督AUC优化算法，并提出基于权重解耦的加速策略以降低算法时间/空间复杂度。进一步地，在优化层面，通过理论分析证明了所提出的算法相对于弱分类器的增加具有指数收敛速率；在模型泛化能力层面，从理论上给出了相比当前工作更为紧致的泛化误差上界。最后，在16个基准数据集上对所提出算法的性能进行了验证，实验结果表明所提出算法在多数情况下以 0.05 显著水平优于其他对比方法，并可在平均意义上产生显著性能提升。

第二，提出了一种多分类AUC优化的算法框架，并对其进行了系统理论分析。基于著名的M度量这一AUC多类扩展指标，进一步通过优化多类AUC度量学习多类评分函数。首先，证明M度量可抵挡数据分布的高阶不平衡特性。基于此，提出一种替代经验风险最小化框架对M度量近似优化。随后，从理论角度证明：（i）对于对率损失、指数损失、平方损失、铰链损失、等损失函数，优

化其可微替代损失足以渐近地获得贝叶斯最优评分函数；（ii）训练框架具有不平衡感知的泛化误差界，与传统的$O(\sqrt{1/N})$结果相比，该框架更加关注来自少数类的瓶颈样本。从实践角度，为提升模型计算效率，提出基于指数损失、平方损失和铰链损失等三种主流替代损失函数的加速方法，降低损失和梯度计算复杂度。最后，通过11个真实数据集上的实验结果证明该框架的有效性，且加速算法可在一定样本量条件下带来10000+倍效率提升。所提加速算法已在阿里部分内容安全场景上线，日均调用频率超过 1 亿次。

最后，提出一种多任务场景下的AUC优化算法框架，将其应用于个性化属性学习问题中。聚焦于个性化属性学习问题，现有属性学习方法主要基于聚合自少量标注者的全局共识。然而，在涉及大量拥有不同兴趣和理解属性词汇方式的标注者时，全局共识并不一定成立。因此，将每个用户的标注预测问题视为不同的任务，从而提出一种多任务学习方法以理解、预测个性化属性标注。与基于共识的属性预测不同，个性化属性学习中偏好学习比标签预测更重要，因此，AUC更适合作为该任务的优化目标。受此启发，提出一种基于用户参数层级化分解的多任务AUC优化方法。聚焦于负迁移问题，提出了一种任务-特征协同学习框架，并应用于该AUC优化方法中。具体地，首先提出一个异构块对角结构正则化算子，该算子实现了特征和任务的协同分组，同时抑制组间知识共享。然后，针对基本模型提出了交替优化算法。理论分析表明，所提出优化算法有如下优点：(a) 具有全局收敛性；(b) 提供块对角结构恢复的保障。

关键词： 机器学习，AUC优化，半监督学习，多分类问题

Abstract

Area Under the ROC Curve (AUC) is a popular evaluation metric that aggregates the performance of a classifier under different thresholds. Meanwhile, AUC could reflect the ranking performance of the classifier since it is equivalent to the possibility that the positive samples are ranked higher than the negative ones. Due to the above-mentioned advantages, AUC is adopted in a wide range of long-tail classification problems such as cyberspace security, medical intelligence, financial analysis， information retrieval, and recommendation system. However, the AUC metric is designed for supervised binary classification problems and not suitable for scenarios in practical applications such as semi-supervised learning, multi-class learning, and multi-task learning, leading to the failure of corresponding AUC optimization. With the rapid growth of data volume in recent years and the increasing complexity and diversification of applications in the field of artificial intelligence, the above-mentioned limitations of existing AUC optimization algorithms are bound to become even more severe. Therefore, it has become a vital problem to design AUC optimization frameworks for different complex scenarios. Consequently, this paper considers the theory, method, and application aspects of AUC optimization for three complexity scenarios (*i.e.*, multi-task, semi-supervised, and multiclass). The main contributions are concluded as follows:

First, a boosting-based semi-supervised AUC optimization method is proposed. Most existing methods only adopt single-model-based methods, while rarely taking into account the benefit of combining multiple models. To address this issue, this paper studies the problem of how to effectively ensemble a series of semi-supervised AUC optimization methods. Specifically, a boosting-based semi-supervised AUC optimization method is proposed. On top of this, an acceleration strategy is provided based on a weight decoupling strategy to reduce the time and space complexity. Moreover, the proposed algorithm is proven to enjoy an exponential convergence rate with respect to the number of weak learners. Finally, a generalization error bound is proposed which is tighter than existing work. Finally, the proposed framework is validated on 16 bench-

mark datasets. Experimental results show that the proposed algorithm outperforms all the competitors with a significance level of 0.05.

Second, a multi-class AUC optimization framework is proposed along with systematic theoretical analysis. The study is based on the well-known M metric for multi-class AUC. First, it is shown that the M metric could withstand the higher-order imbalance problem lead by a multi-class long-tail distribution. Motivated by this, an empirical surrogate risk minimization framework is proposed to approximately optimize the M metric. Theoretically, it is shown that: (i) logit loss, exponential loss, squared loss, hinge loss are all consistent with the M metric; (ii) the training framework enjoys an imbalance-aware generalization error bound, which pays more attention to the bottleneck samples of minority classes compared with the traditional $O(\sqrt{1/N})$ result. Practically, to deal with the low scalability of the computational operations, we propose acceleration methods for three popular surrogate loss functions, including the exponential loss, squared loss, and hinge loss, to reduce the computational complexity of loss and gradient. Finally, experimental results on 11 real-world datasets demonstrate the effectiveness of the proposed framework, and the acceleration algorithm can bring 10000+ times efficiency improvement under certain sample size conditions. The developed technologies have been applied to content security systems in Alibaba with a daily call frequency as high as 1 billion.

Finally, a novel AUC optimization framework is proposed in this paper for the personalized attribute learning problem. We first focus on the personalized attribute learning problem. Most existing attribute learning methods are trained based on the consensus of annotations aggregated from a limited number of annotators. However, the consensus might fail in settings, especially when a wide spectrum of annotators with different interests and comprehension about the attribute words are involved. Therefore, the attribute preference learning problem for each annotator is regarded as a specific task. On top of this, a multi-task AUC optimization method is further proposed to predict personalized attribute annotations. Different from consensus attribute learning, preference prediction is more crucial than label prediction, where AUC is a suitable performance measure. Motivated by this, a multi-task AUC optimize is proposed on

top of a hierarchical decomposition mechanism for user parameters. Focusing on the negative transfer problem, a task-feature collaborative learning framework is further proposed and applied to the aforementioned AUC optimization method. Specifically, a heterogeneous block-diagonal structure regularizer is proposed to leverage the collaborative grouping of features and tasks and suppressing inter-group knowledge sharing. Then, an alternating optimization method is proposed to train the model. Last but not least, theoretical analysis shows that the proposed method has the following benefits: (a) it enjoys the global convergence property and (b) it provides a block-diagonal structure recovery guarantee.

Keywords: Machine Learning, AUC Optimization, Semi-Supervised Learning, Multi-class Problem

目 录

图形列表

表格列表

第1章　引言

1.1　研究背景

随着大数据及深度学习的逐步成熟，人工智能技术取得了突破性进展，成为信息技术时代不可或缺的关键技术之一。正如国务院在《国务院关于印发新一代人工智能发展规划的通知》所指出的：“人工智能的迅速发展将深刻改变人类社会生活、改变世界”。由此可见人工智能技术在我国未来发展中具有重要地位。

作为现代人工智能技术的核心之一，机器学习在近二十年来广泛受到学术界关注，其中最为著名的深度学习技术更在2012年AlexNet获得Imagenet冠军以来相继在计算机视觉、自然语言处理等领域相继取得重大突破，在工业界得到广泛普及。国家战略层面，上文提及的《国务院关于印发新一代人工智能发展规划的通知》也将高级机器学习理论作为新一代人工智能基础理论体系的重要组成部分。

由于机器学习/深度学习技术本身采用数据驱动模式进行知识获取，其成功无法脱离对实际问题中的数据依赖。大数据时代，网络空间中的数据呈现出长尾态势，高价值数据淹没于大量低价值数据中，难以被有效挖掘。主流深度学习/机器学习方法主要面向数据平衡分布的数据集，忽略了不同类别样本之间的潜在分布差异，无法有效适用于不同类别极端不平衡的长尾分布数据集。近年来，海量复杂数据集不断涌现而出，数据中的长尾分布特点愈发凸显，如何在长尾分布条件下构建有效的深度学习/机器学习方法，已成为人工智能领域面临的共性技术挑战之一。

造成这一挑战的主要根源之一在于：主流方法往往采用最小化总体错误率指标这一理念进行模型及算法设计，而总体错误率本身对数据分布较为敏感，较易忽略样本中稀有数据类别的性能。相比之下，AUC （Area Under the ROC Curve），即ROC曲线下面积，可度量分类器在不同分类阈值下的平均性能，或或等效表示为正例得分高于负例的概率，对类别分布、错分代价均不敏感，在长尾数据分布场景下更适合作为评价指标(Fawcett, 2006b; Hand 等, 2001)。例如，针对网络空间安全监管问题，风险样本在总体样本中占比极低（万分之一

乃至十万分之一），通过错误率评估模型时，倾向于将所有样本分类为安全样本以提高准确率，故往往难以挖掘风险样本。得益于对类别分布、错分代价均不敏感，AUC等价于提高风险样本评分高于安全样本的概率，优化结果与风险数据的数量无关，因此更适用于风险数据检测场景。类似地，医疗领域中病变样本远少于正常样本，金融领域违约用户远少于守约用户，均可利用AUC优化方法以获得更符合期望的分类结果。另一方面，在推荐系统问题中，用户感兴趣的商品数据显著小于总商品量，无法通过准确率刻画模型性能，而AUC则可通过衡量用户交互商品和未交互商品的排序性能，规避长尾效应对模型评估的影响。AUC的良好性质使其成为本世纪机器学习理论及方法的重要研究内容，与ROC/AUC相关的两条词条(Area Under Curve, ROC analysis) 已被收录于由ICML02/04大会主席 Claude Sammut 等人编著，图灵奖得主Geofrey Hinton、著名嵌入算法ISOMAP提出者、ICML12/16大会主席John Langford、机器学习领域著名期刊Machine Learning Journal前主编(2010-2020) Peter A. Flach 等人参与撰写的机器学习百科全书中(Sammut 等, 2011)。

由于AUC指标的计算复杂度远高于错误率，那么能否在最小化错误率框架下实现最大化AUC呢？著名算法Support Vector Network (Support Vector Machine) 提出者之一，纽约Google Research负责人 Corinna Cortes在其早期工作(Cortes 等, 2003)中指出根据最小化错误率得出的模型可能在AUC指标意义下为次优模型，并由此确定了AUC优化框架的必要性。目前学界内对于AUC优化的相关算法及理论已具有了一定积累。然而，AUC指标针对有监督的二分类问题设计，无法适用于半监督学习、多类别分类学习、以及多任务学习等实际应用中更为常见的复杂场景。由于AUC指标本身的特性，研究复杂场景下的AUC优化问题除需应对指标本身的局限性，还必然面临以下算法及理论层面的共性挑战：

1. **算法层面**：AUC优化目标函数引入正负样本逐对损失函数，损失与梯度的计算复杂度大致为样本量平方规模，难以适用于基于mini-batch的随机优化算法。

2. **理论层面**：在AUC指标所诱导的替代损失函数中，求和项之间往往不满足独立假设，传统学习理论中的泛化能力分析工具无法适用，且在深度学习框架下的AUC泛化理论在学界仍未有工作涉及。

随着近年来随着人工智能领域数据体量的飞速增长及人工智能领域应用的

日益复杂化、多样化，现有AUC优化算法的上述局限性势必愈发凸显。从理论层面出发，如何构建面向复杂场景、复杂需求的AUC优化算法框架成为亟需解决的科学问题之一，可为机器学习基础理论提供新鲜血液。从实践层面出发，AUC优化面向长尾数据下的分类问题，是网络空间安全分析、生物信息学、智能医疗等重大应用领域所面对的共性难题，而本文工作可促进AUC优化理论及方法在复杂场景下的普及，具有重要的应用赋能价值。

本章后续内容中首先进一步介绍与本文工作密切相关的现有研究基础，在此基础上给出目前待解决的关键问题及本文研究内容，最后给出论文的组织形式即各章节之间的联系。

1.2　理论基础及研究现状

1.2.1　ROC及AUC指标

操作特性曲线(Receiver Operating Characteristic curve, ROC)最早出现于信号处理的研究工作中(Egan, 1975)，随后因其统计上的优良性质于上世纪末本世纪初逐渐被机器学习领域学者关注(Bradley, 1997; Woods 等, 1997; Bowyer 等, 1999; Mozer 等, 2001; Hand 等, 2001; Cortes 等, 2003)。本节将依次介绍ROC曲线、AUC指标的严格数学定义、数学性质，以及近年来面向复杂问题的AUC拓展指标研究现状。

下面给出ROC及AUC的严格定义。首先给出数据分布的基本定义。给定训练样例$(\boldsymbol{x}, y)$，其输入特征$\boldsymbol{x}$为d维欧式空间中向量，即$\boldsymbol{x} \in \mathbb{R}^d$，其类标$y$为1或-1，即$y \in \{-1, 1\}$。若$y = 1$，称该样例为**正例**；若$y = -1$，则称该样例为**负例**。此外，记$\mathcal{P}$为正例分布、$\mathcal{N}$为负例分布。在此基础上给出两个必要指标：伪阳性率(False Positive Rate, FPR)及真阳性率(True Positive Rate)的定义。具体地，给定分类阈值t及分类函数h，通常将所有满足$h(\boldsymbol{x}) > t$的样本预测为正例，在此意义下$TPR_h(t)$为正例$\boldsymbol{x}_+$的预测准确率，$FPR_h(t)$为错将负例$\boldsymbol{x}_-$预测为正例的概率，数学上可分别表示为：

$$\begin{aligned} TPR_h(t) &= \mathop{\mathbb{P}}_{\boldsymbol{x}_+}\left[h(\boldsymbol{x}_+) > t\right], \\ FPR_h(t) &= \mathop{\mathbb{P}}_{\boldsymbol{x}_-}\left[h(\boldsymbol{x}_-) > t\right]. \end{aligned} \tag{1.1}$$

根据上述定义，操作特性曲线(Receiver Operating Characteristic curve, ROC)则被定义为不同分类阈值t下，以$TPR_h(t)$为纵轴，$FPR_h(t)$横轴绘制而成的图像。

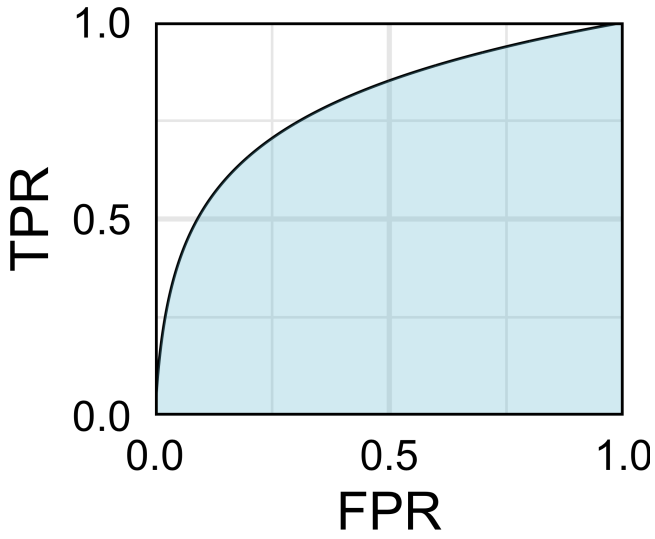


图 1.1 ROC曲线及AUC

Figure 1.1 ROC curve and AUC

ROC曲线下夹面积(Area Under the roc Curve, AUC)即可表示为该曲线与$x=0, x=1, y=0$围成的面积，数学上可表为：

$$AUC(f)=\int_0^1 TPR_h\left(FPR_h^{-1}(t)\right)dt \tag{1.2}$$

下面对AUC、ROC的性质进行进一步介绍。首先考虑一种更为直观的AUC等价形式化。具体来说，(Hanley 等, 1982) 指出，AUC指标等价于正例样本根据h获得的得分高于负例样本获得的得分的概率，根据此结论，可将得分函数$h(\cdot)$对应的AUC指标表为：

$$AUC(h)=1-\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{N}}\left[\ell_{0-1}\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right], \tag{1.3}$$

其中，$f(\boldsymbol{x},\boldsymbol{x}')$为正负例分差即 $f(\boldsymbol{x},\boldsymbol{x}')=h(\boldsymbol{x})-h(\boldsymbol{x}')$；$\ell_{0-1}$为0-1损失函数，即$\ell_{0-1}(x)=I[x<0]$。当正负例排序错误时正例得分低于负例，$\ell_{0-1}(f(\boldsymbol{x},\boldsymbol{x}'))=1$；反之排序正确，有$\ell_{0-1}(f(\boldsymbol{x},\boldsymbol{x}'))=0$。由此不难得出AUC数值即为正例负例得分排序的准确率，因此AUC对得分排序性能也有很好的刻画能力，在推荐、检索、二分排序等问题中被广泛采纳。除此之外，(Fawcett, 2006a)等工作还指出，AUC对得分函数的取值大小、类别分布、错分代价等因素均有良好的稳健性，相比于传统的总体错误率指标更适合作为类不平衡/长尾数据分布下的分类器度量指标。

正如机器学习百科全书中(Sammut 等, 2011)所述，如何度量多类别问题中的ROC、AUC性能是该领域中的一大开放问题。伴随着上世纪末AUC/ROC被引入机器学习领域，学术界陆续涌现出一批关于多分类AUC指标的探索工作。此类相关中作通过两类方式构造多类别下的AUC度量。第一种方式认为二类ROC曲线的多类拓展可表示为高维曲面。因此，可将AUC自然地推广为ROC曲面下体积（VUS）(Mossman, 1999; Ferri 等, 2003)。然而，高维空间体积的计算复杂度极高，计算N个样本和N_C类VU 的时间复杂度可达$O(N\log N + N^{\lfloor N_C/2 \rfloor})$，空间复杂度可达$O(N^{\lfloor N_C/2 \rfloor})$。另一种方式则更加简单，直接取多对二分类AUC (Hand 等, 2001; Provost 等, 2003; Honzik 等, 2009; Yang, 2009) 的平均值以定义多类AUC。该方法的核心思想是：如果每一类别对的分布能够很好地分离，模型即可达到较好性能。由于该策略摆脱高维空间的计算，从而大大降低计算复杂度。由于其简单性，代表性工作(Hand 等, 2001)中提出的M度量已经被许多机器学习软件所采用，如`python`中的`sklearn`和`R`中的`pROC`等。最近，(Wang 等, 2020)进一步给出了多类AUC度量的在线扩展，以解决流式数据中的概念漂移问题。

除多类别问题上的拓展，学术界近期也涌现出部分其他维度上的AUC指标拓展工作。(Dodd 等, 2003; Walter, 2005; Yang 等, 2019a)对ROC曲线下部分TPR，FPR取值范围构成的面积进行了理论与实证研究；(Maurer 等, 2020)则进一步给出了加权AUC的理论性质；(Jaskowiak 等, 2020) 进一步给出了聚类场景下AUC的拓展定义。

与纯粹的指标分析研究不同，本文主要关注如何在AUC引导下构造ERM问题并完成模型优化，下文将进一步对AUC优化中的基础理论及相关工作进行详细论述。

1.2.2　有监督二分类AUC优化

经验风险最小化(Empirical Risk Minimization, ERM) 框架是机器学习的核心范式。该框架旨在通过最小化训练集上的特定损失函数获得理想的模型。而损失函数的选择依赖于决策者所选择的模型性能度量指标，不同度量指标可诱导出不同的损失函数进而推演出不同的经验风险最小化问题。传统的机器学习方法主要基于最小化错误率（最大化准确率）指标进行算法设计。不同于此类方法，本文主要聚焦于AUC诱导的ERM问题。鉴于此，本节介绍有监督二分类条件下由AUC诱导的经验风险最小化(Empirical Risk Minimization,ERM) 相关研究

工作。首先给出ERM的基本范式，随之介绍近二十年来AUC优化方面的主要成果，最后给出AUC优化学习理论方面的研究进展。

回顾式 (1.3)中内容，AUC正比于分类器进行正负例正确排序的概率，而由于ERM框架主要考虑最小化问题，因此首先将AUC转化为损失形式。具体地，定义期望风险：

$$R_{PN}(h) = 1 - AUC(h) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{N}}\left[\ell_{0-1}\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right]. \tag{1.4}$$

由此，最大化$AUC(h)$即可与最小化$R_{PN}(h)$一一对应。由于目标函数中设计离散的0-1函数ℓ_{0-1}，因此$R_{PN}(h)$的最小化问题显然为一组合优化问题，ERM框架中引入可微代理损失 (surrogate loss function) ℓ 代替$0-1$ 损失函数。此时得到替代期望风险函数：

$$R^{\ell}_{PN}(h) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{N}}\left[\ell\left(f(\boldsymbol{x},\boldsymbol{x}')\right)\right]\right]. \tag{1.5}$$

进一步，由于$R^{\ell}_{PN}(h)$计算需要对于分布$\mathcal{P},\mathcal{N}$求取期望，而实际应用中样本期望一般不可能获得，因此需进一步通过给定训练样本上的统计量对代理期望风险进行估计。具体地，令采样过程中样例输入及类标的联合分布为$p(\boldsymbol{x},y)$，则通过以下采样过程获得训练样本：

$$\begin{aligned}\mathcal{X}_P &= \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{i.i.d}{\sim} p_P(\boldsymbol{x}) = \mathbb{P}\left[\boldsymbol{x}|y=1\right],\\ \mathcal{X}_N &= \{\boldsymbol{x}'_k\}_{k=1}^{n_n} \overset{i.i.d}{\sim} p_P(\boldsymbol{x}') = \mathbb{P}\left[\boldsymbol{x}'|y=-1\right],\end{aligned} \tag{1.6}$$

其中n_p,n_n分别为正例及负例样本个数；p_P,p_N 分别表示正例及负例样本的条件分布。在此基础上通过训练集上的均值估计总体期望，得到替代经验风险损失：

$$\hat{R}^{\ell}_{PN}(h) = \frac{1}{n_p n_n}\sum_{\boldsymbol{x}\in\mathcal{X}_P}\sum_{\boldsymbol{x}'\in\mathcal{X}_N}\ell\left(f(\boldsymbol{x},\,\boldsymbol{x}')\right). \tag{1.7}$$

此时即可获得AUC优化对应的近似优化问题：

$$\operatorname*{argmin}_{h\in\mathcal{H}}\hat{R}^{\ell}_{PN}(h), \tag{1.8}$$

其中$\mathcal{H}$ 为所选定的假设空间，由模型类型（决策树、神经网络、线性模型等），以及正则化项决定。由于模型通常由参数确定，设模型h的参数为$\boldsymbol{w}$，并记此时模型为$h_{\boldsymbol{w}}$，并设参数选自集合$\mathcal{W}$，则可将式 (1.8)等效转化为式 (1.9):

$$\operatorname*{argmin}_{\boldsymbol{w}\in\mathcal{W}}\hat{R}^{\ell}_{PN}(h_{\boldsymbol{w}}), \tag{1.9}$$

对于AUC优化动机的理论依据主要来自于早期研究(Cortes 等, 2003)，该工作指出虽然在期望意义下AUC可大致反比于错误率，但由于在不均衡分布或较大错误率下最小化错误率得到的AUC方差显著，因此在有限样本情况下最大化AUC无法通过最小化错误率替代。在此项研究发表后，AUC优化研究开始活跃于机器学习领域。

在本世纪初，AUC优化工作主要聚焦于不同的ERM及建模方式。(Alan 等, 2004; Calders 等, 2007; Yan 等, 2003) 基于分类文中最为常见的logit损失 $\ell(t) = \log(1+\exp(-t))$构建了直接AUC优化问题。(Freund 等, 2003a)则在Boosting框架下给出了将AUC优化形式化为RankBoost问题进行模型集成，根据boosting理论的基本性质易知该模型可以有效最小化基于指数损失$\ell(t) = \exp(-t)$的AUC替代经验风险。Joachims等人将AUC优化问题视为结构化输出 (Strucutured Outputs)问题的一个特例，并基于StructSVM 框架实现了该优化问题(Joachims, 2005, 2006)，由SVM的基本性质，此类方法对应拓展的hinge替代损失函数$\ell(t) = \max(1-t, 0)$的最小化。

2010年后AUC优化方法的相关工作主要聚焦于AUC优化问题的效率提升及在复杂问题中的拓展。首先，为适应大数据分析，研究人员开始探索AUC在线优化方法。(Zhao 等, 2011)通过缓冲池取样技术进行这一方向的初步尝试。(Gao 等, 2013)提供一种基于平方代理损失的流数据单次AUC优化方法以适应流式数据中的AUC优化问题。最近，(Ying 等, 2016)将基于平方损失的随机AUC最大化问题转化为一个随机鞍点问题。该鞍点问题的目标函数只涉及实例损失项的求和，从而大幅减少成对优化算法的计算负担。(Natole 等, 2018, 2019)进一步优化了该框架的收敛速度。考虑到AUC面积无法聚焦于ROC曲线对应的局部区域性能，(Narasimhan 等, 2013b,c) 结合StructSVM中的割平面算法(Cutting Plane Method) 对特定FPR区间下的偏AUC (Partial AUC)进行了直接优化。(Shen 等, 2020) 进一步考虑了可拒识的AUC优化方法。

1.2.3　AUC优化框架学习理论研究

除AUC优化的方法研究之外，不少研究工作也对AUC优化框架相关的学习理论问题进行了系统研究。主要从两个方向进行研究，第一个方向主要研究由引入不同替代损失导致的系统误差，即不同替代损失的fisher一致性问题。记最小化真实期望风险函数$R_{PN}(h)$得到的贝叶斯最后分类器为$f^{\star}_{AUC}$，记最小化替

代期望风险函数$R^{\ell}_{PN}(h)$得到的分类器为$f^{\star}_{\ell}$，则称替代损失函数ℓ与AUC指标是(fisher)一致的，当且仅当对于任意函数列$\{h^k\}_k$以及任意数据分布:

$$h^k \to f^{\star}_{\ell} \Longrightarrow h^k \to f^{\star}. \tag{1.10}$$

直观而言，式 (1.10)优化fisher一致的损失函数可在渐近意义下得到贝叶斯最优分类器(Mohri 等, 2018)。在此研究方向中作为著名的研究当属(Gao 等, 2015)，该工作首先将错误率一致性分析中的主要工具——广义校正(generalized calibration)条件引入AUC最优化框架中，给出了该条件在AUC指标下的拓展，进一步证明该条件是一致性的必要非充分条件，最后给出了AUC一致性的一个简单充分条件。(Gao 等, 2013)则进一步巧妙地证明了平方损失的一致性。

第二个研究方向主要研究由引入训练集估计而导致的系统误差，即不同算法的泛化误差误差。在泛化误差分析中，主要关注期望损失$R^{\ell}_{PN}(h)$与经验损失$\hat{R}^{\ell}_{PN}(h)$之间的差异上界随训练样本量增大收敛于0的速率。该速率越快则说明ERM可在越少样本情况下完成期望风险的优化。进行AUC优化的泛化分析的挑战在于求和项之间往往不满足独立假设，传统学习理论中基于Rademacher复杂度的理论框架(Mohri 等, 2018)无法直接适用。鉴于此，(Agarwal 等, 2005)通过拓展VC维的定义对AUC的泛化性能解进行了系统分析。具体地若，得分函数h限定于某假设空间$\mathcal{H}$，则在该空间最大的泛化偏差$|R^{\ell}_{PN}(h) - \hat{R}^{\ell}_{PN}(h)|$以大概率满足以下上界：

$$\sup_{h\in\mathcal{H}} \left|R^{\ell}_{PN}(h) - \hat{R}^{\ell}_{PN}(h)\right| \le O\left(\sqrt{\left(\frac{n_p+n_n}{n_p n_n}\cdot \mathsf{Comp}(\mathcal{H})\right)}\right). \tag{1.11}$$

其中$\mathsf{Comp}(\mathcal{H})$为由VC维诱导的假设空间$\mathcal{H}$复杂度度量，具体细节可见原文。考虑到VC维仅适用于二值输出的分类器限制了假设空间的灵活性，(Clémençon 等, 2008)进一步通过拓展对称化技术给出了基于Rademacher复杂度度量的AUC替代损失泛化性能上界。(Usunier 等, 2006; Ralaivola 等, 2010)则基于依赖图及其染色问题将AUC估计中的非独立项分解为若干独立项，建立了基于染色Rademacher复杂度并进一步完善了AUC优化问题的泛化性能界。(Usunier 等, 2005)指出AUC指标为U统计量(Korolyuk 等, 2013)的特例，并通过U统计量的性质构造了在一定条件下更为紧致的上界。最近的一项工作(Maurer 等, 2019)则给出了弱交互函数(weak interactive functions)的AUC泛化界。

概括而言，在二分类方向的AUC方法及理论已经趋于成熟。但由于AUC指标自身定义的限制，其在半监督、多分类等复杂应用方面的拓展研究均处于早期阶段。下面分别对半监督、多分类相关研究进行简要论述并总结其局限性。

1.2.4　半监督AUC优化

相比于全监督条件下的AUC研究，半监督条件下的AUC优化研究尚处于早期阶段。文献(Sakai 等, 2018; Xie 等, 2018a)对离线情况下的半监督AUC优化进行了系统研究，文献(Xie 等, 2018b)对在线条件下的半监督AUC优化进行了系统研究。由于本文主要考虑离线条件下的AUC优化，因此下面针对(Sakai 等, 2018; Xie 等, 2018a)进行进一步介绍。

相比于全监督情况，半监督条件下数据集中还存在未标注数据$\mathcal{X}_U$，其生成过程如下:

$$\mathcal{X}_U = \{\tilde{x}_j\}_{j=1}^{n_u} \overset{i.i.d}{\sim} p(\boldsymbol{x}) = \theta_P \cdot p_P(\boldsymbol{x}) + \theta_N \cdot p_N(\boldsymbol{x}), \tag{1.12}$$

其中θ_p, θ_N分别表示$\mathbb{P}[y=1], \mathbb{P}[y=-1]$即正类和负类的先验概率。显然，引入未标注样集$\mathcal{X}_U$使期望损失$R_{0-1}^{PN}$无法直接计算。以下讨论如何在半监督条件下估计真实期望风险R_{0-1}^{PN}。首先构建辅助风险损失函数R_{0-1}^{PU}与R_{0-1}^{UN}如下：

$$\begin{aligned} R_{0-1}^{PU} &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{U}}\left[\ell_{0-1}\left(f\left(\boldsymbol{x},\tilde{\boldsymbol{x}}\right)\right)\right]\right], \\ R_{0-1}^{UN} &= \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{U}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{N}}\left[\ell_{0-1}\left(f\left(\tilde{\boldsymbol{x}},\boldsymbol{x}'\right)\right)\right]\right]. \end{aligned} \tag{1.13}$$

R_{0-1}^{PU}与R_{0-1}^{UN}分别表示正例得分低于未标注样例的概率，以及未标注样例得分低于负例的概率。(Sakai 等, 2018)证明，在可对θ_P, θ_N进行较好估计的前提下，可由$R_{0-1}^{PU}, R_{0-1}^{UN}$估计出全监督损失$R_{0-1}^{PN}$（同理替代风险间也满足该关系）。(Xie 等, 2018a)进一步证明，即便在无法估计θ_P, θ_N条件下，R_{0-1}^{PN}亦可分别由R_{0-1}^{PU}及R_{0-1}^{UN}线性表示，其数学形式可表为：

$$R_{0-1}^{PN} = \frac{1}{\theta_n} R_{0-1}^{PU} - \frac{1}{2} \cdot \frac{\theta_p}{\theta_n}, \tag{1.14}$$

$$R_{0-1}^{PN} = \frac{1}{\theta_p} R_{0-1}^{UN} - \frac{1}{2} \cdot \frac{\theta_n}{\theta_p}. \tag{1.15}$$

该结论表明，R_{0-1}^{PN}数值同时正比于R_{0-1}^{PU}及R_{0-1}^{UN}，因此，可在未知类别分布先验θ_p, θ_n的情况下通过优化R_{0-1}^{PU}、R_{0-1}^{UN}优化R_{0-1}^{PN}。

在(Sakai 等, 2018; Xie 等, 2018a)提出的框架中，虽可给出R_{0-1}^{PU}、R_{0-1}^{UN}及R_{0-1}^{PN}之间的线性关系，但该关系仅在期望意义上成立。而在实际半监督问题中，已

标注样本通常十分有限，且AUC优化ERM的样本复杂度通常高于基于最小化错误率的方法，因此在样本层面模型的泛化性能仍然较为有限。

1.2.4.1 多类别AUC优化

另一个与本文密切相关的问题是多分排序，该问题是二分排序问题的一个自然扩展。在多分排序中，偏序关系可以表示为两个以上的离散值。目前，已有许多工作关注多分排序问题的AUC优化方法和理论 (Uematsu 等, 2014; Gao 等, 2018; Clémençon 等, 2013, 2017; Rajaram 等, 2005)。然而，只有各类存在语义上的次序关系时，多分排序方法才能解决多分类问题。如，年龄估计任务可以被视为多分排序问题，类标签是人的年龄；电影分级预测也可以被视为多分排序问题，类标签是电影的分级。与多分排序问题不同，本文关注一般性多分类问题。在一般性的多分类问题中，不同的标签表示不同的语义概念，并无明确顺序关系。因此，多分排序的相关方法不能适用于本文的问题设定。

1.3 本文工作

1.3.1 研究内容

面向复杂场景下的AUC优化问题，本文聚焦于如下关键问题：

• 现有半监督AUC优化算法主要考虑单一模型的优化问题，泛化能力有限，无法适用于标注严重不足的情况。相比之下，学习论研究 (Schapire 等, 1998) 指出boosting算法由于可隐式优化margin函数，可在集成多个弱分类器的条件下不易产生过拟合现象，更加适用于所面临的场景。而关于如何在半监督AUC的复杂设定下构建boosting算法的研究相对较为空白。因此如何在boosting框架下构建更为高效、泛化能力更强的半监督AUC优化方法是本文所需面对的一大关键问题。

• 机器学习领域著名期刊Machine Learning Journal前主编(2010-2020) Peter A. Flach 等人在机器学习百科全书中指出(Sammut 等, 2011)ROC分析的一大开放问题来自于其多分类拓展。一方面，由于ROC/AUC仅面向二分类问题有明确定义，如何进行多分类ROC/AUC分析这一问题本身就具有极大的挑战性。另一方面，由二分类到多分类问题，AUC计算复杂度将显著增加，算法设计和分析难度显著加大。相比之下，现有的少数与多分类AUC优化相关的研究工作往往仅考虑类别之间有层级比较关系的情况，无法适用于一般意义上的多分类问题。

因此，如何在一般的多分类问题意义下建立高效的AUC优化方法以及系统的理论分析，是AUC优化领域亟待解决的另一个关键问题。

• 随着机器学习领域面对问题的日益复杂化，机器学习模型往往需要同时面对多个任务，传统的单任务学习范式已无法再满足对学习效率需求，多任务联合学习范式则正在成为一种普遍化需求。而AUC优化在此场景下的研究工作相对较为空白，如何在此场景下构造AUC优化算法有待系统研究。另外，多任务学习领域著名学者杨强在其综述 (Zhang 等, 2017) 中指出when to share是多任务的三大关键挑战之一。更为具体的，当无关任务之间的模型共享会使模型更易过拟合产生所谓的**负迁移问题**(Kang 等, 2011)。因此，如何显式规避负迁移是进行多任务AUC优化所必然面对的关键挑战。

1.3.2　主要贡献

本文主要贡献总结如下：

• 针对半监督场景，提出一种基于boosting的无先验半监督AUC优化模型集成方法。首先，就AUC带来的额外计算复杂度设计了高效的加速算法，使更新单个弱分类器的时间复杂度由平方规模降至线性规模。在算法收敛速率方面，证明训练集误差随弱分类器个数增加以接近几何速率快速收敛。在算法泛化误差性能方面进行了系统的理论分析，首先根据半监督AUC优化目标函数自身特点构造半监督AUC优化的Rademacher复杂度；其次，针对该复杂度特性提出一种广义最大值不等式；最终给出在模型集成意义下的半监督AUC优化泛化误差上界，所得结果较现有结论(Sakai 等, 2018)更为紧致。

• 针对多分类场景，提出多类别AUC优化框架并进行理论分析。首先对比了不同的多分类AUC拓展指标，并采纳(Hand 等, 2001)中提出的M度量作为多分类AUC指标。在此基础上构建了对应的替代经验风险最小化问题，并就一致性和泛化能力方面展开了系统的理论分析。一致性方面，首次证明指数损失、logit损失、平方损失、铰链损失等损失函数相对于多分类AUC指标在特定假设下均具有一致性；泛化能力方面，基于Rademacher复杂度及覆盖数对深度全连接网络及深度卷积网络在多分类AUC优化框架下的泛化性能上界进行了系统分析。最后，针对多分类AUC优化的可拓展性进行了系统的改进，为铰链损失、平方损失以及指数损失提供了高效的损失及梯度计算的加速算法。所提加速算法已在阿里部分内容安全场景上线，日均调用频率超过 1 亿次。

• 针对多任务场景，提出面向个性化属性预测问题的多任务AUC优化方法。聚焦于个性化属性标注预测问题，将每个用户的标注预测问题视为不同的任务，形成多任务模型。进一步基于主流用户共识、用户群体聚类和个性化三级要素提供多任务模型参数的层级化分解表达形式；在此基础上构造对应的AUC优化问题，采用近端梯度下降法求解模型参数，为群体要素的近端算子推导出闭式解，并进一步设计一种基于AUC的加速计算方法；最后，对方法收敛性和泛化能力进行系统的理论分析，并在模拟数据集和真实属性标注数据中验证所提方法的有效性。

• 进一步针对多任务中的负迁移问题，提出一种任务-特征协同学习框架。通过异构二分图的谱图论性质，提出一种可促进任务-特征协同分组的块对角谱正则约束，并形成基础模型。在该模型中，知识共享仅存在于分组内，而在组间的知识迁移则被块对角结构自动阻断，由此实现负迁移的抑制。并根据该正则化项形成任务-特征协同学习的非凸非光滑目标函数。在此基础上提出一种类近端梯度优化方法交替求解模型中的参数。进一步通过理论分析证明本文所提出的优化算法可在保障所求非凸问题全局收敛性质的同时保障块对角结构的有效恢复。最后，将所提出框架应用于个性化属性学习任务中通过所提出的任务-特征协同学习框架构造了多任务AUC优化问题。

1.3.3 组织结构

全文剩余部分组织如图1.2所示，下面给出简要总结：

• 第2章研究boosting理论下的半监督AUC优化方法。通过权重解耦策略，提出了高效的boosting算法，其单次迭代具有线性复杂度。在此基础上，对所提出算法的收敛速率及泛化性能进行了理论分析。在17个真实数据集上进行实验分析，结果证明所提方法能有效提升半监督条件下模型的AUC性能。

• 第3章对多分类场景下的AUC优化展开了系列研究。提出了多分类AUC优化的经验风险最小框架，并针对不同替代损失设计计算加速算法。从理论角度出发，对一致性及泛化能力展开进一步分析。在11个真实数据集上进行实验分析，结果证明所提方法能有效提升多分类条件下模型的AUC性能。

• 第4章以个性化属性学习为应用，研究多任务场景下的AUC优化方法拓展。将用户标注为行为建模为多任务问题，提出任务参数的层级模型因子分解，联合AUC优化损失形成目标函数。在此基础上，构造优化方法并研究器收敛性

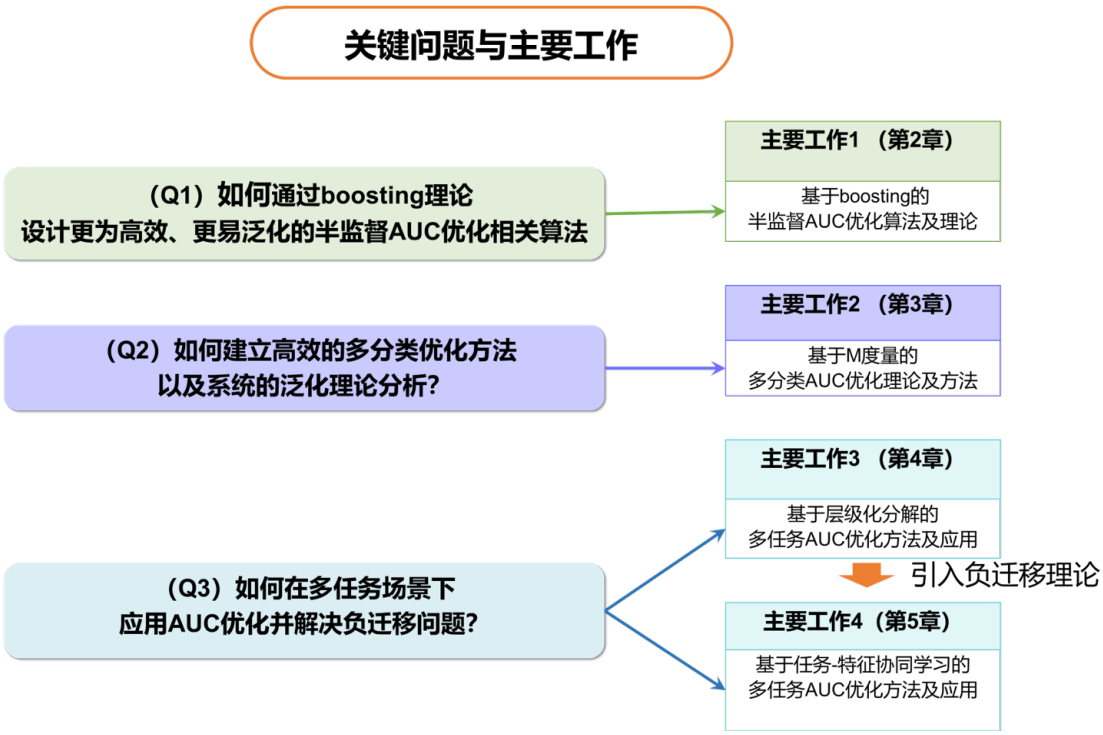


图 1.2 全文组织结构

Figure 1.2 Organization of this paper

质。最终在sun与shoes 两个个性化属性学习数据集上验证模型性能，所提出算法较现有算法可产生显著性能及效率提升。

• 第5章在第4章基础上研究多任务的负迁移理论。提出特征-任务协同学习框架，通过块对角约束抑制不相关任务-特征间的迁移。在此基础上研究优化算法的收敛性质及结构恢复保障。最后将所提算法应用于个性化属性学习问题中。最终在仿真数据集及sun与shoes 两个个性化属性数据集上对所提出算法性能进行了验证。

• 第6章对全文进行总结，并就多任务、多分类、以及局部面积优化问题对未来工作进行初步展望。

第2章　基于Boosting的半监督AUC优化理论及方法

2.1　引言

在早期的机器学习研究中，往往采用最小化错误率的理念设计模型和优化算法。那么能否在最小化错误率的框架下实现最大化AUC的目的呢？(Cortes 等, 2003)指出根据最小化错误率得出的模型可能在AUC指标意义下为次优模型，因此有必要直接针对AUC指标设计优化方法。在此项工作之后的近二十年内，涌现出了大批AUC优化的相关研究(Alan 等, 2004; Calders 等, 2007; Freund 等, 2003a; Joachims, 2005, 2006; Ying 等, 2016; Natole 等, 2018; Agarwal 等, 2005; Clémençon 等, 2008; Usunier 等, 2005, 2006; Ralaivola 等, 2010; Lyu 等, 2018; Gao 等, 2013; Agarwal, 2014; Gao 等, 2015)。

绝大多数的AUC优化相关研究局限于处理数据标注全部已知的情况，无法适用于数据中存在未标注样本的半监督学习场景。在近期的工作中，已有部分研究聚焦于半监督AUC优化问题。(Fujino 等, 2016)通过对未标注数据的分布函数建模构造了生成模型并由此构造了半监督AUC优化方法。(Sakai 等, 2018)首次基于PU (Positive Unlabeled) 学习框架推导出一个AUC的无偏估计，并结合有监督AUC和PU学习提出一种半监督AUC学习框架。该框架无需预生成未标记样本的伪标签，但需预先估计样本正负类的先验概率以对未标记样本进行加权，在已标记样本数量较小的情况下仍然存在局限性。(Xie 等, 2018a)进一步指出在0-1损失意义下，无需任何先验分布信息也可由未标记样本估计AUC风险。(Xie 等, 2018b)在文献(Xie 等, 2018a)的基础上进一步实现了半监督AUC的随机优化方法。

目前，半监督AUC优化已取得了初步的成功。但现有半监督AUC优化方法仅针对单个线性模型进行设计，对更为复杂的模型则缺乏考虑。鉴于此，本章主要研究基于模型集成的半监督AUC优化算法并对其性质进行了系统的理论分析，以期通过融合多个弱学习器突破现有方法的瓶颈。具体而言，本章的主要贡献如下：

1. 提出一种基于boosting的无先验半监督AUC优化模型集成方法，并就其效率瓶颈设计了高效的加速算法，大幅降低了更新单个弱分类器的时间/空间复

杂度。

2. 在算法收敛速率方面，证明训练集误差随弱分类器的增加以指数速度迅速衰减。

3. 在算法泛化误差性能方面进行了系统的理论分析，首先根据半监督AUC优化目标函数自身特点构造了半监督AUC优化的Rademacher复杂度；其次，针对该复杂度特性提出一种广义最大值不等式；最终给出了首个在模型集成意义下的半监督AUC优化泛化误差上界，并获得了比现有半监督AUC优化泛化界(Sakai 等, 2018)更为紧致的结果。

本章在16个标准数据集进行了系统的实验分析，实验结果表明本章算法在绝大多数情况下，可在0.05显著水平下优于其他对比方法。

2.2 相关工作

在介绍本章主要工作前，本章先就所涉及的既往工作及其必要技术细节进行简要回顾，2.2.1 小节进一步延伸至半监督AUC优化方法(Sakai 等, 2018; Xie 等, 2018a)，随后在2.2.2小节中回顾本章所基于的模型集成框架RankBoost(Freund 等, 2003b)。注意有监督AUC优化已在引言中提及，此处不再赘述，下文将沿用引言中所用的符号。

2.2.1 半监督AUC优化

相比与全监督条件下的AUC研究，半监督条件下的AUC优化研究尚处于早期阶段。文献(Sakai 等, 2018; Xie 等, 2018a)对离线情况下的半监督AUC优化进行了系统研究，文献(Xie 等, 2018b)对在线条件下的半监督AUC优化进行了系统研究。由于本章主要考虑离线条件下的AUC优化，因此下面针对(Sakai 等, 2018; Xie 等, 2018a)进行进一步介绍。

相比于全监督情况，半监督条件下数据集中还存在未标注数据$\mathcal{X}_U$，其生成过程如下:

$$\mathcal{X}_U = \{\tilde{x}_j\}_{j=1}^{n_u} \overset{i.i.d}{\sim} p(\boldsymbol{x}) = \theta_P \cdot p_P(\boldsymbol{x}) + \theta_N \cdot p_N(\boldsymbol{x}), \tag{2.1}$$

其中θ_p， θ_N分别表示$\mathbb{P}[y=1], \mathbb{P}[y=-1]$即正类和负类的先验概率。显然，引入未标注样集$\mathcal{X}_U$使期望损失$R_{0-1}^{PN}$无法直接计算。下面讨论如何在半监督条件下

估计真实期望风险R_{0-1}^{PN}。首先构建辅助风险损失函数R_{0-1}^{PU}与R_{0-1}^{UN}如下：

$$\begin{aligned} R_{0-1}^{PU} &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{U}}\left[\ell_{0-1}\left(f\left(\boldsymbol{x},\tilde{\boldsymbol{x}}\right)\right)\right]\right] \\ R_{0-1}^{UN} &= \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathcal{U}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{N}}\left[\ell_{0-1}\left(f\left(\tilde{\boldsymbol{x}},\boldsymbol{x}'\right)\right)\right]\right] \end{aligned} \tag{2.2}$$

R_{0-1}^{PU}与R_{0-1}^{UN}分别表示正例得分低于未标注样例的概率，以及未标注样例得分低于负例的概率。(Sakai 等, 2018)证明，在可对θ_P,θ_N进行较好估计的前提下，可由$R_{0-1}^{PU},R_{0-1}^{UN}$估计出全监督损失$R_{0-1}^{PN}$（同理替代风险间也满足该关系）。(Xie 等, 2018a)进一步证明，即便在无法估计θ_P,θ_N条件下，R_{0-1}^{PN}亦可分别由R_{0-1}^{PU}及R_{0-1}^{UN}线性表示，其数学形式可表为：

$$R_{0-1}^{PN} = \frac{1}{\theta_n}R_{0-1}^{PU} - \frac{1}{2}\cdot\frac{\theta_p}{\theta_n}, \tag{2.3}$$

$$R_{0-1}^{PN} = \frac{1}{\theta_p}R_{0-1}^{UN} - \frac{1}{2}\cdot\frac{\theta_n}{\theta_p}. \tag{2.4}$$

该结论表明，R_{0-1}^{PN}数值同时正比于R_{0-1}^{PU}及R_{0-1}^{UN}，因此可在未知类别分布先验θ_p,θ_n的情况下通过优化R_{0-1}^{PU}、R_{0-1}^{UN}优化R_{0-1}^{PN}。同理于有监督情况下的AUC优化近似问题，(Xie 等, 2018a) 采用平方损失$\ell_{sq}(t)=(1-t)^2$作为替代损失构造如下替代风险函数：

$$\hat{R}_{PNU}^{\ell_{sq}} = \gamma\cdot\hat{R}_{PN}^{\ell_{sq}}(h) \tag{2.5}$$

$$+(1-\gamma)\cdot\left(\hat{R}_{PU}^{\ell_{sq}}(h)+\hat{R}_{UN}^{\ell_{sq}}(h)-\frac{1}{2}\right), \tag{2.6}$$

$$\tag{2.7}$$

其中：

$$\hat{R}_{PN}^{\ell_{sq}}(h) = \frac{1}{n_p n_n}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\ell_{sq}\left(f\left(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)}\right)\right) \tag{2.8}$$

$$\hat{R}_{PU}^{\ell_{sq}}(h) = \frac{1}{n_p n_u}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\ell_{sq}\left(f\left(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_j^{(u)}\right)\right) \tag{2.9}$$

$$\hat{R}_{UN}^{\ell_{sq}}(h) = \frac{1}{n_u n_n}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\ell_{sq}\left(f\left(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)}\right)\right) \tag{2.10}$$

在此基础上，进一步求解以下问题，即可得到(Xie 等, 2018a)所提出的SAMULT算法：

$$\operatorname*{argmin}_{h\in\mathcal{H}}\hat{R}_{PNU}^{\ell_{sq}}(h), \tag{2.11}$$

由于平方损失的基本性质，该问题存在显式闭式解(closed-form solution)，具体细节可见(Xie 等, 2018a)，本章不再赘述。

2.2.2 RankBoost算法的一般化框架

本小节介绍本章所采用的RankBoost算法的一般化框架。考虑一般的逐对排序学习问题，给定逐对比较数据集$\mathcal{X}_{pair}$，对所有$(x_1, x_2) \in \mathcal{X}_{pair}$，希望得分$f(x_1)$应尽可能大于$f(x_2)$。RankBoost近似最小化以下形式的指数排序损失函数实现该目标：

$$\sum_{(x_1,x_2)\in\mathcal{X}_{pair}} D^0(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot \exp\left(-\left(f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\right)\right) \tag{2.12}$$

为实现模型集成，RankBoost以迭代形式逐步学习多个**弱学习器**$h^1, \cdots, h^T$，及其权重$\alpha^1, \cdots, \alpha^T$，并以弱学习器的加权投票形式作为其输出$f$，即$f(x) = \sum_{i=1}^{T} \alpha^t h^t(x)$。RankBoost的一般框架见算法 1。

算法 1 RankBoost模型一般化过程

输入: 模型输入$\boldsymbol{X}, \boldsymbol{Y}$，弱分类器个数$T$

输出: 弱分类器权重 $\alpha^1, \cdots, \alpha^T$，弱分类器$h^1, \cdots, h^T$

初始化样本权重D^0:

while $t \le T$ **do**

Step 1: 根据权重D_t，求解弱分类器h^t获得其性能ϵ^t

Step 2: 根据当前弱分类器性能ϵ^t，更新α^t

Step 3: 根据α^t, h^t计算归一化因子$\tilde{Z}^t$:

$$\tilde{Z}^t = \sum_{(x_1,x_2)\in\mathcal{X}_{pair}} D^t(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot \exp\left(-\alpha^t \cdot \left(h^t(\boldsymbol{x}_1) - h^t(\boldsymbol{x}_2)\right)\right) \tag{2.13}$$

Step 4: 计算新一轮样本权重D^t:

$$D^{t+1}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{D^t(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot \exp\left(-\alpha^t \cdot \left(h^t(\boldsymbol{x}_1) - h^t(\boldsymbol{x}_2)\right)\right)}{\tilde{Z}^t} \tag{2.14}$$

$t = t + 1$

end while

生成最终模型 $f(\boldsymbol{x}) = \sum_{i=1}^{T} \alpha^t \cdot h^t(\boldsymbol{x})$

与adaboost算法相同(Freund 等, 1996)，RankBoost算法在每次迭代中根据上一轮迭代产生的样本权重D_t确定一个新**弱学习器**h^t以及该学习器对应的模型权重α^t。随后，RankBoost算法基于h^t及α^t为每个排序样例对$(\boldsymbol{x}_1, \boldsymbol{x}_2)$确定下一次迭代的样本权重$D^{t+1}(\boldsymbol{x}_1, \boldsymbol{x}_2)$。$D^{t+1}(\boldsymbol{x}_1, \boldsymbol{x}_2)$，数值越大则该样例对对于下轮迭代模型越为重要。由$D^{t+1}(\boldsymbol{x}_1, \boldsymbol{x}_2)$更新过程可知，$h^t(\boldsymbol{x}_1) - h^t(\boldsymbol{x}_2) > 0$ 模型输出与期

望序关系一致，此时对应的权重较小；反之，若$h^t(\boldsymbol{x}_1) - h^t(\boldsymbol{x}_2) < 0$ 模型输出与期望序关系不一致，$D^{t+1}(\boldsymbol{x}_1, \boldsymbol{x}_2)$则有增大趋势。因此，随着迭代不断进行，RankBoost模型将逐步聚焦于难以正确排序的样例对，形成由易至难的自适应学习过程。

2.3 方法形式化

本章中针对式 (2.15)及RankBoost算法设计高效的半监督AUC优化模型集成算法。沿用2.2.1中提及的理论完成半监督条件下的AUC替代损失估计。由于本章采用RankBoost进行模型集成，因此将采用指数替代损失函数$\ell_{\exp}(t) = \exp(-t)$。记$\hat{R}_{PN}^{\ell_{exp}}, \hat{R}_{PNU}^{\ell_{exp}}, \hat{R}_{UN}^{\ell_{exp}}$为将式 (2.8)、式 (2.9)、式 (2.10)中的ℓ_{sq}替换为$\ell_{\exp}$所得到的经验风险函数，给定模型假设空间$\mathcal{H}$，本章求解由指数替代损失诱导出的AUC优化问题：

$$\min_{f \in \mathcal{H}} \left(\gamma \cdot \hat{R}_{PN}^{\ell_{exp}}(f) + \frac{(1-\gamma)}{2} \cdot \left(\hat{R}_{PU}^{\ell_{exp}}(f) + \hat{R}_{UN}^{\ell_{exp}}(f) \right) \right) \cdot \exp(\rho \cdot \sum_{t=1}^{T} \alpha^t) \tag{2.15}$$

其中$\exp(\rho \cdot \sum_{t=1}^{T} \alpha^t)$为正则项，其作用可见定理 2.3 中的分析。

对于本章的目标函数而言，直接采用RankBoost会造成$O(|n_p n_u + n_p n_n + n_u n_n|)$的时间及空间复杂度，基本正比于训练样本量规模的平方，计算效率较低。鉴于此，本章将针对 式 (2.15) 设计高效的模型集成方法，其细节见算法 2。

2.3.1 PNUAUCBoost算法设计

样本权重解耦: 首先根据式 (2.15)的形式，构造过程 0 step4中D^t的解耦方法，使空间复杂度由$\mathcal{O}(n_p \cdot n_u + n_u \cdot n_n)$降至 $\mathcal{O}(n_p + n_u + n_n)$。注意到式 (2.15)中仅存在三类样例对，分别为$(\boldsymbol{x}', \boldsymbol{x}_j^{(u)})$、$(\boldsymbol{x}_j^{(u)}, \boldsymbol{x}_k^{(n)})$及$(\boldsymbol{x}', \boldsymbol{x}_k^{(n)})$。为同时将这三类样例对的权重解耦，构造辅助权重

$$\begin{aligned} &\{\omega_{p,i}^{+,t}\}_{i=1}^{n_p},\ \{\omega_{u,j}^{-,t}\}_{j=1}^{n_u},\ \{\omega_{u,j}^{+,t}\}_{j=1}^{n_u}, \\ &\{\omega_{n,k}^{-,t}\}_{k=1}^{n_n}, \{\nu_{p,j}^{+,t}\}_{j=1}^{n_p},\ \{\nu_{n,k}^{-,t}\}_{k=1}^{n_n} \end{aligned} \tag{2.16}$$

并将D^t表示为：

$$D^t(\boldsymbol{x}', \boldsymbol{x}_j^{(u)}) = \omega_{p,i}^{+,t} \cdot \omega_{u,j}^{-,t}, \quad i = 1, \cdots, n_p,\ j = 1, \cdots, n_u, \tag{2.17}$$

$$D^t(\boldsymbol{x}_j^{(u)}, \boldsymbol{x}_k^{(n)}) = \omega_{u,j}^{+,t} \cdot \omega_{n,k}^{-,t}, \quad j = 1, \cdots, n_u,\ k = 1, \cdots, n_n, \tag{2.18}$$

$$D^t(\boldsymbol{x}', \boldsymbol{x}_k^{(n)}) = \nu_{p,j}^{+,t} \cdot \nu_{n,k}^{-,t}, \quad i = 1, \cdots, n_p,\ k = 1, \cdots, n_n, \tag{2.19}$$

基于上述解耦方式，进一步设计算法0的具体实现细节。

权重初始计算：将初始值设为:

$$\begin{aligned}\omega_{p,i}^{+,0}=\frac{C_1}{n_p},\quad \omega_{u,j}^{+,0}=\frac{C_1}{n_u},\quad \omega_{u,j}^{-,0}=\frac{C_1}{n_u},\\ \omega_{n,k}^{-,0}=\frac{C_1}{n_n},\quad \nu_{p,j}^{+,0}=\frac{C_2}{n_p},\quad \nu_{n,k}^{-,0}=\frac{C_2}{n_n}\end{aligned} \tag{2.20}$$

其中 $C_1=\left(\frac{1-\gamma}{2}\right)^{1/2}$，$C_2=(\gamma)^{1/2}$。此时，对应的 $D^0(\boldsymbol{x}',\boldsymbol{x}_k^{(n)})$ 恰为 $\frac{\gamma}{n_p n_n}$，$D^0(\boldsymbol{x}',\boldsymbol{x}_j^{(u)})$，$D^0(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})$ 恰为 $\frac{1-\gamma}{2n_p n_u}$，$\frac{1-\gamma}{2n_u n_n}$。记$R_{fin}$ 为式 (2.15)的目标函数，有:

$$\begin{aligned}R_{fin}=&\Bigg(\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}D^0(\boldsymbol{x}',\boldsymbol{x}_j^{(u)})\exp\left((f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}')\right)\\&+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^0(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})\exp\left(f(\boldsymbol{x}_k^{(n)})-f(\boldsymbol{x}_j^{(u)})\right)\\&+\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^0(\boldsymbol{x}',\boldsymbol{x}_k^{(n)})\exp\left(f(\boldsymbol{x}_k^{(n)})-f(\boldsymbol{x}')\right)\Bigg)\\&\cdot\exp(\rho\cdot\sum_{i=1}^{T}\alpha^t)\end{aligned} \tag{2.21}$$

因此上述初始权重设置方法可将算法中的样本权重引入目标函数中。

归一化因子的高效计算：进一步根据式 (2.17)式 (2.19)中的解耦规则，实现并加速过程 0 中式 (2.13)的计算。对于$\hat{R}_{PN}^{\ell_{exp}}$相关计算，有：

$$\begin{aligned}&\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^t(\boldsymbol{x}',\boldsymbol{x}_k^{(n)})\exp\left(-\alpha^t(h^t(\boldsymbol{x}')-h^t(\boldsymbol{x}_k^{(n)}))\right)\\&\qquad=\left(\sum_{i=1}^{n_p}\nu_{p,j}^{+,t}\cdot\exp\left(-\alpha^t\cdot h^t(\boldsymbol{x}')\right)\right)\cdot\left(\sum_{k=1}^{n_n}\nu_{n,k}^{-,t}\cdot\exp\left(\alpha^t\cdot h^t(\boldsymbol{x}_k^{(n)})\right)\right)\end{aligned} \tag{2.22}$$

对于$\hat{R}_{PU}^{\ell_{exp}}$及$\hat{R}_{UN}^{\ell_{exp}}$，易得类似形式的分解。鉴于此，可将归一化因子的计算转化为样例归一化因子的乘积之和。为简化数学表达，于下文中采用以下简写符号：

$$h_{p,i}^t=h^t(\boldsymbol{x}'),\ h_{u,j}^t=h^t(\boldsymbol{x}_j^{(u)}),\ h_{n,k}^t=h^t(\boldsymbol{x}_k^{(n)}) \tag{2.23}$$

则该分解过程可表为：

$$Z_1^t = \sum_{i=1}^{n_p} \omega_{p,i}^{+,\,t-1} \cdot \exp(-\alpha h_{p,i}^t), \tag{2.24}$$

$$Z_2^t = \sum_{j=1}^{n_u} \omega_{u,j}^{-,\,t-1} \cdot \exp(\alpha h_{u,j}^t), \tag{2.25}$$

$$Z_3^t = \sum_{j=1}^{n_u} \omega_{u,j}^{+,\,t-1} \cdot \exp(-\alpha h_{u,j}^t), \tag{2.26}$$

$$Z_4^t = \sum_{k=1}^{n_n} \omega_{n,k}^{-,\,t-1} \cdot \exp(\alpha h_{n,k}^t), \tag{2.27}$$

$$Z_5^t = \sum_{i=1}^{n_p} \nu_{p,i}^{+,\,t-1} \cdot \exp(-\alpha h_{p,i}^t), \tag{2.28}$$

$$Z_6^t = \sum_{k=1}^{n_n} \nu_{n,k}^{-,\,t-1} \cdot \exp(\alpha h_{n,k}^t), \tag{2.29}$$

$$\tilde{Z}^t = Z_1^t \cdot Z_2^t + Z_3^t \cdot Z_4^t + Z_5^t \cdot Z_6^t \tag{2.30}$$

样本权重更新： 与归一化因子$\tilde{Z}^t$的化简相似，可根据式 (2.17)-式 (2.19)中的解耦将算法0式 (2.14)中的权重更新过程转化为对解耦权重的更新过程：

$$\omega_{p,i}^{+,\,t} = \frac{\omega_{p,i}^{+,\,t-1} \exp(-\alpha h_{p,i}^t)}{\sqrt{\tilde{Z}^t}} \tag{2.31}$$

$$\omega_{u,j}^{-,\,t} = \frac{\omega_{u,j}^{-,\,t-1} \exp(\alpha h_{u,j}^t)}{\sqrt{\tilde{Z}^t}} \tag{2.32}$$

$$\omega_{u,j}^{+,\,t} = \frac{\omega_{u,j}^{+,\,t-1} \exp(-\alpha h_{u,j}^t)}{\sqrt{\tilde{Z}^t}} \tag{2.33}$$

$$\omega_{n,k}^{-,\,t} = \frac{\omega_{n,k}^{-,\,t-1} \exp(\alpha h_{n,k}^t)}{\sqrt{\tilde{Z}^t}} \tag{2.34}$$

$$\nu_{p,j}^{+,\,t} = \frac{\nu_{p,i}^{+,\,t-1} \exp(-\alpha h_{p,i}^t)}{\sqrt{\tilde{Z}^t}} \tag{2.35}$$

$$\nu_{n,k}^{-,\,t} = \frac{\nu_{n,k}^{-,\,t-1} \exp(\alpha h_{n,k}^t)}{\sqrt{\tilde{Z}^t}} \tag{2.36}$$

模型权重α^t的更新： 首先通过以下引理给出损失函数的一个上界：

引理 2.1. 当算法T次迭代结束后，有：

$$R_{fin} = \prod_{t=1}^{T} \left(\exp(\rho\alpha^t) \cdot \tilde{Z}^t\right) \tag{2.37}$$

证明. 见附录 A.2 □

由以上引理，若在第t次迭代使$\exp(\rho\alpha^t)\cdot\tilde{Z}^t$尽可能小，则最终将获得较为理想的目标函数值。因此基于近似最小化$\exp(\rho\alpha^t)\cdot\tilde{Z}^t$的原则设计$\alpha^t, h^t$的更新方式。考虑如下放缩，利用exp的凸性及Jensen不等式，$\forall x\in[-1,1]$有:

$$\exp(\alpha\cdot x)\leq\frac{1+x}{2}\cdot\exp(\alpha)+\frac{1-x}{2}\cdot\exp(-\alpha), \tag{2.38}$$

将弱分类器h^t的输出值域限制于区间$[0,1]$内，根据$\tilde{Z}^t$定义及式 (2.38)，可得：

$$\begin{aligned}\tilde{Z}^t \leq & \sum_{i=1}^{n_p}\sum_{j=1}^{n_u} D^t(\boldsymbol{x}', \boldsymbol{x}_j^{(u)})\,\tilde{\psi}_{i,j} + \sum_{j=1}^{n_u}\sum_{k=1}^{n_n} D^t(\boldsymbol{x}_j^{(u)}, \boldsymbol{x}_k^{(n)})\,\tilde{\psi}_{j,k} \\ & \sum_{i=1}^{n_p}\sum_{k=1}^{n_n} D^t(\boldsymbol{x}', \boldsymbol{x}_k^{(n)})\,\tilde{\psi}_{i,k} \\ \triangleq & C(\alpha^t)\end{aligned} \tag{2.39}$$

其中：

$$\psi_{i,j} = \frac{1+h_{p,i}^t - h_{u,j}^t}{2}\cdot\exp(-\alpha) + \frac{1+h_{u,j}^t - h_{p,i}^t}{2}\cdot\exp(\alpha) \tag{2.40}$$

$$\tilde{\psi}_{j,k} = \frac{1+h_{u,j}^t - h_{n,k}^t}{2}\cdot\exp(-\alpha) + \frac{1+h_{n,k}^t - h_{u,j}^t}{2}\cdot\exp(\alpha) \tag{2.41}$$

$$\tilde{\psi}_{i,k} = \frac{1+h_{p,i}^t - h_{n,k}^t}{2}\cdot\exp(-\alpha) + \frac{1+h_{n,k}^t - h_{p,i}^t}{2}\cdot\exp(\alpha) \tag{2.42}$$

根据上式对$\tilde{Z}^t$的放缩，固定弱分类器模型h^t，并求解α^t使$\exp(\rho\cdot\alpha^t)C(\alpha^t)$最小化。为便于数学上的表达，记：

$$\begin{aligned}\Delta^t = & \sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{D^t(\boldsymbol{x}', \boldsymbol{x}_j^{(u)})}{2}\cdot(h_{p,i}^t - h_{u,j}^t) \\ & + \sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{D^t(\boldsymbol{x}_j^{(u)}, \boldsymbol{x}_k^{(n)})}{2}\cdot(h_{u,j}^t - h_{n,k}^t) \\ & + \sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{D^t(\boldsymbol{x}', \boldsymbol{x}_k^{(n)})}{2}\cdot(h_{p,i}^t - h_{n,k}^t)\end{aligned} \tag{2.43}$$

易证$\exp(\rho\cdot\alpha^t)\cdot C(\alpha^t)$为关于$\alpha^t$的凸函数，因此为最小化该因变量仅需求解：

$$\frac{d\left[\exp(\rho\cdot\alpha^t)C(\alpha^t)\right]}{d\alpha^t} = 0 \tag{2.44}$$

可解得α^t为：

$$\alpha^t = \frac{1}{2}\log\left(\frac{1+\Delta^t}{1-\Delta^t}\right) - \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right). \tag{2.45}$$

Δ^t的高效计算：在计算α^t过程中需要计算Δ^t进而需要遍历所有的样例对权重$\boldsymbol{D}^t$。因此Δ^t的计算过程也是本算法的主要计算瓶颈之一。鉴于上文中提出的权重解耦方法，同样可以给出Δ^t的加速计算方式。由式 (2.17) - 式 (2.19)，有：

$$\Delta^t = \sum_{i=1}^{n_p} g_{p,i}^t \cdot h_{p,i}^t + \sum_{j=1}^{n_u} g_{u,j}^t \cdot h_{u,j}^t + \sum_{k=1}^{n_n} g_{n,k}^t \cdot h_{n,k}^t \tag{2.46}$$

其中：

$$g_{p,i}^t = \omega_{p,i}^{+,t} \cdot \Big(\sum_{j=1}^{n_u} \omega_{u,j}^{-,t} \Big) + \nu_{p,j}^{+,t} \cdot \Big(\sum_{k=1}^{n_n} \nu_{n,k}^{-,t} \Big) \tag{2.47}$$

$$g_{u,j}^t = \omega_{u,j}^{+,t} \cdot \Big(\sum_{k=1}^{n_n} \omega_{n,k}^{-,t} \Big) - \omega_{u,j}^{-,t} \cdot \Big(\sum_{i=1}^{n_p} \omega_{p,i}^{+,t} \Big) \tag{2.48}$$

$$g_{n,k}^t = -\omega_{n,k}^{-,t} \cdot \Big(\sum_{j=1}^{n_u} \omega_{u,j}^{+,t} \Big) - \nu_{n,k}^{-,t} \cdot \Big(\sum_{k=1}^{n_n} \nu_{p,j}^{+,t} \Big) \tag{2.49}$$

注意到上式中所有的权重求和项仅需计算一次，因此仅需$\mathcal{O}(N)$时间复杂度即可完成所有$g_{p,i}^t, g_{u,j}^t, g_{n,k}^t$的计算。为便于后续计算，记$\boldsymbol{G}_t \in \mathbb{R}^{(n_p+n_u+n_n)\times 1}$为所有$g_{p,i}^t, g_{u,j}^t, g_{n,k}^t$拼接而成的列向量，并记$\boldsymbol{H}_t \in \mathbb{R}^{(n_p+n_u+n_n)\times 1}$为所有$h_{p,i}^t, h_{u,j}^t, h_{n,k}^t$拼接而成的列向量。根据符号$\boldsymbol{G}^t, \boldsymbol{H}^t$，可将$\Delta^t$进一步简写为：

$$\Delta^t = (\boldsymbol{G}^t)^\top \boldsymbol{H}^t \tag{2.50}$$

弱分类器模型h^t的更新：给定α^t, Δ^t，进一步反推出弱分类器h^t的一个合理更新规则。为此，构建以下引理：

引理 2.2. 对于第t次算法迭代，若由式 (2.45)更新α^t，有

$$\exp(\rho\alpha^t) \cdot \tilde{Z}^t \leq \exp(\rho\alpha^t) \cdot C(\alpha^t) \tag{2.51}$$

$$= \exp\left(-KL\left(\frac{1+\rho}{2} \| \frac{1+\Delta^t}{2} \right) \right) \tag{2.52}$$

其中KL为二元相对熵，其定义为：

$$KL(p||q) = p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right), \quad \forall p \in [0,1], q \in [0,1] \tag{2.53}$$

证明. 见附录 A.3。 □

由引理 2.2可知，若通过式 (2.45)更新α^t，则可通过最大化

$$KL\left((1+\rho)/2 || (1+\Delta^t)/2\right)$$

近似实现$\tilde{Z}^t$的最小化。与RankBoost相同，本章采用决策树桩(Freund 等, 2003a)（Decision Stump）学习弱分类器，并将文献(Freund 等, 2003a)中的弱分类器目标函数替换为$KL\left((1+\rho)/2||(1+\Delta^t)/2\right)$。给定样本输入$\boldsymbol{x} = [x_1, x_2 \cdots, x_d]^\top$，选定特征维度$e$，及阈值$\theta$，决策树桩函数输出如下：

$$h_\theta^e(x) = \begin{cases} 1, & \text{if } x^e > \theta \\ 0, & \text{otherwise} \end{cases} \tag{2.54}$$

将决策树桩的输出函数带入式 (2.50)，得其对应的Δ^t可表为：

$$\Delta^t = \sum_{x_i^e > \theta} g_i^t \tag{2.55}$$

其中i, θ 待学习超参数，i为决策树桩选定的参数决策的输入维度，θ为其选定阈值。由上式可知，决策树桩输出为1，仅当给定样例的第i维输入数值大于阈值θ。为求解决策树桩的两个参数，为输入特征的每一维度i设定阈值备选集$\boldsymbol{T}_i$，并搜索能够最大化$KL\left((1+\rho)/2||(1+\Delta^t)/2\right)$的$\theta, i$，最终生成弱分类器$h^t$。算法细节可见附录 A.1 中的算法 8。

综合上述所有细节，得到高效的PNU-AUC的Boosting算法，并将其汇总于算法 2。

算法 2 PNUAUCBoost

输入: 模型输入 $\boldsymbol{X}$，超参数 $\gamma \in [0,1]$, 弱分类器个数T，超参数ρ。

输出: 弱分类器权重 $\alpha^1, \cdots, \alpha^T$，弱分类器 $h^1, \cdots, h^T$

通过式 (2.20)初始化权重

while $t \leq T$ **do**

根据式 (2.47)-式 (2.49)完成$\boldsymbol{G}^t$计算

根据式 (2.47)-式 (2.49)完成$\boldsymbol{H}^t$计算

根据算法 8 获得弱分类器 h^t 并获得其对应的Δ^t

更新α^t:

$$\alpha^t = \frac{1}{2}\log\left(\frac{1+\Delta^t}{1-\Delta^t}\right) - \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) \tag{2.56}$$

由式 (2.24)-式 (2.30) 计算归一化因子$\tilde{Z}^t$

由式 (2.31)-式 (2.36) 更新权重

$t = t + 1$

end while

生成最终模型 $f(\boldsymbol{x}) = \sum_{i=1}^{T} \alpha^t \cdot h^t(\boldsymbol{x})$

2.4　理论分析

2.4.1　收敛分析

本小节给出算法 2 的收敛速率。首先，给出相对熵的关键数学性质：

引理 2.3. $\forall p \in (0,1), q \in (0,1)$ 有：

$$KL\,(p||q) \geq 2 \cdot (p-q)^2 \tag{2.57}$$

证明. 将(Popescu 等, 2016)中的定理1.3应用于二项分布即可获得本引理。 □

根据引理2.1-2.3，有以下结论：

定理 2.1. 给定$\rho > 0$，设算法 2 在T轮迭代结束后所得到的分类器为$f(x) = \sum_{i=1}^{T} \alpha^t \cdot h^t(x), \alpha^t \geq 0, \forall t \in T$ [1]，定义r_ρ为：

$$\begin{aligned} r_\rho(f) =& \frac{\gamma}{n_p n_n} \sum_{i=1}^{n_p} \sum_{k=1}^{n_n} I\left[\frac{f(\boldsymbol{x}_k^{(n)}) - f(\boldsymbol{x}')}{\sum_{i=1}^{T}\alpha^t} \leq \rho\right] \\ &+ \frac{1-\gamma}{2n_p n_u} \sum_{i=1}^{n_p} \sum_{j=1}^{n_u} I\left[\frac{f(\boldsymbol{x}_j^{(u)}) - f(\boldsymbol{x}')}{\sum_{i=1}^{T}\alpha^t} \leq \rho\right] \\ &+ \frac{1-\gamma}{2n_u n_n} \sum_{j=1}^{n_u} \sum_{k=1}^{n_n} I\left[\frac{f(\boldsymbol{x}_k^{(n)}) - f(\boldsymbol{x}_j^{(u)})}{\sum_{i=1}^{T}\alpha^t} \leq \rho\right] \end{aligned} \tag{2.58}$$

[1] $\alpha^t \geq 0$可通过设置较小的ρ或早停机制实现

有如下结论：

(a) 对于一般情况：

$$r_\rho(f) \leq \exp\left(-\sum_{i=1}^{T} KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)\right) \tag{2.59}$$

(b) 若进一步假设$\min_t \Delta^t \geq \delta > \rho$，有：

$$r_\rho(f) \leq \exp\left(-\frac{T}{2}\cdot(\delta-\rho)^2\right) \tag{2.60}$$

证明. (a)可由 $r_\rho \leq R_{fin}$及引理 2.1-2.2 推知；在结论(a)基础上进一步应用引理 2.3 即可得结论(b)。 □

上述定理表明r_ρ数值随着迭代进行大致呈指数衰减趋势。另一方面，由于当$\rho > 0$时，$r_\rho > r_0$，因此模型的训练误差r_0 也呈指数衰减趋势。由此可知，所提算法具有较快的收敛速率。

2.4.2 泛化性能分析

本小结中，继续讨论算法 2 的泛化性能。首先，将泛化性能定义为样本分布下的期望AUC错误率，即R_{0-1}^{PN}。在此基础上，希望对于任意由算法 2 产生的学习器，下式以大概率成立:

$$R_{0-1}^{PN} \leq \epsilon_1 + \epsilon_2 \tag{2.61}$$

其中ϵ_1与训练集上的模型排序错误率有关，ϵ_2与算法产生模型的丰富程度以及训练集的样本量有关。因此，当训练样本充分且模型优化充分的条件下保证上式成立即可保证泛化误差R_{0-1}^{PN}较小。本章将通过学习论(Mohri 等, 2018)中基于Rademacher复杂度的分析方式完成上式的证明。其证明思路关键在于通过对称化技术（见文献(Mohri 等, 2018)定理 3.3 ）引入假设空间的Rademacher复杂度。然而经典的对称化技术仅适用于损失函数的求和项之间相互独立的情况。进一步考察半监督AUC损失的定义形式。不难发现涉及样例对$(\boldsymbol{x}', \boldsymbol{x}_j^{(u)})$的损失项与所有涉及$\boldsymbol{x}'$或$\boldsymbol{x}_j^{(u)}$的损失项都无法独立，同理可知，样例对$(\boldsymbol{x}', \boldsymbol{x}_k^{(n)})$及$(\boldsymbol{x}_j^{(u)}, \boldsymbol{x}_k^{(n)})$同样存在该问题。鉴于此，本章通过引理A.8 （见附录 A.4.6）给出了适用于半监督AUC的对称化技术，并引入如下适用于PNU-AUC问题的Rademacher复杂度：

定义 2.1 (**PNU-AUC Rademacher 复杂度**). 给定数据集 $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$及假设集$\mathcal{H}$，损失函数$\ell$ PNUAUC Rademacher 经验复杂度由下式给出：

$$\begin{aligned}&\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell \circ \mathcal{H})\\&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}} \sum_{i=1}^{N_C} \sum_{j\neq i} \sum_{\boldsymbol{x}_m\in\mathcal{N}_i} \sum_{\boldsymbol{x}_n\in\mathcal{N}_j} T^{i,j,m,n}\right],\end{aligned} \tag{2.62}$$

其中

$$Q_{i,k} = \frac{\sigma_i^{(p)} + \sigma_j^{(u)}}{2} \cdot \frac{\gamma}{n_p n_n} \cdot \ell(f, \boldsymbol{x}_i^{(p)}, \boldsymbol{x}_k^{(n)}) \tag{2.63}$$

$$Q_{j,k} = \frac{\sigma_j^{(u)} + \sigma_k^{(n)}}{2} \cdot \frac{1-\gamma}{2n_u n_n} \cdot \ell(f, \boldsymbol{x}_j^{(u)}, \boldsymbol{x}_k^{(n)}) \tag{2.64}$$

$$Q_{i,j} = \frac{\sigma_i^{(p)} + \sigma_j^{(u)}}{2} \cdot \frac{1-\gamma}{2n_p n_u} \cdot \ell(f, \boldsymbol{x}_i^{(p)}, \boldsymbol{x}_j^{(u)}) \tag{2.65}$$

变量 $\sigma_i^{(p)},\ \forall i = 1, 2, \cdots, n_p,\ \sigma_j^{(u)},\ \forall j = 1, 2, \cdots, n_u,\ \sigma_k^{(n)},\ \forall k = 1, 2, \cdots, n_n$ 为独立同分布Rademacher 随机变量[2]。在经验复杂度基础上，定义总体水平的 PNU-AUC Rademacher 复杂度为其经验复杂度对$\mathcal{S}$的期望，表为:

$$\mathfrak{R}_{PNU}(\ell \circ \mathcal{H}) = \mathbb{E}_{\mathcal{S}}\left[\hat{\mathfrak{R}}_{PNU,\mathcal{S}}(\ell \circ \mathcal{H})\right]$$

进一步，给出决策树桩弱分类器的假设空间定义：

定义 2.2 (**决策树桩的假设空间**). 将决策树桩的假设空间定义为：

$$\mathcal{H}_{DS} = \left\{h_\theta^e : e \in [d], \theta \in \boldsymbol{T}^e\right\} \tag{2.66}$$

其中 $h_\theta^e(\boldsymbol{x}) = I\left[\boldsymbol{x}^e > \theta\right]$，$[d] = \{1, 2, 3, \cdots, d\}$, $\boldsymbol{T}^e$为第e个维度特征的备选阈值集合，对于每个维度的阈值候选集，固定其阈值备选集$\boldsymbol{T}^i$，并固定备选阈值个数为K。

在证明过程中，需要利用假设空间的凸包及其Rademacher复杂度的数学性质，其定义如下：

定义 2.3 (假设空间的凸包). 给定任意假设集$\mathcal{H}$，定义其凸包为$co(\mathcal{H})$为如下函数集：

$$\begin{aligned}co(\mathcal{H}) = \Big\{&h : \exists T \in \mathbb{N},\ s.t.\ h = \sum_{i=1}^{T} \alpha^t \cdot h^t,\ h^t \in \mathcal{H},\\&\alpha^t \geq 0\ \forall t \in [T],\ \sum_{i=1}^{T} \alpha^t = 1\Big\}\end{aligned} \tag{2.67}$$

[2]Rademacher 随机变量σ 在 $\{-1, 1\}$内随机取值且$\mathbb{P}[\sigma = 1] = \mathbb{P}[\sigma = -1] = \frac{1}{2}$

进一步构造辅助函数ℓ_ρ

$$\ell_\rho(x) = \min\left(1, \max\left(0, 1 - \frac{x}{\rho}\right)\right) \tag{2.68}$$

根据文献(Mohri 等, 2018)，有：

$$\ell_{0-1}(x) \leq \ell_\rho(x) \leq I\left[x \leq \rho\right] \leq \exp(-x + \rho) \tag{2.69}$$

由上式可知，辅助损失函数可作为连接r_ρ及r_0的关键桥梁，在附录 A.4 的证明中将频繁使用该函数。

基于上述定义、不等式以及 A.4 中的引理 B.1-引理 A.8，首先给出$\mathcal{H}_{DS}$凸包中函数的泛化界：

定理 2.2 (**PNU-AUCBOOST的泛化界**). 给定训练数据集$\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$、$\rho > 0$，及弱分类器的阈值备选集$\boldsymbol{T}_e$，设样例均由独立采样生成，则对于任意函数$f \in co\left(\mathcal{H}_{DS}\right)$以及任意$\delta \in (0,1)$，下式至少以$1-\delta$概率成立:

$$\begin{aligned} R_{0-1}^{PN}(f) \leq & \left(r_\rho(f) + \frac{8\sqrt{2}}{\rho}\left((\log d + \log K) \cdot \chi(\boldsymbol{Y})\right)^{1/2}\right. \\ & \left. + \frac{1}{\rho}\left(\log(\frac{2}{\delta}) \cdot \chi(\boldsymbol{Y})\right)^{1/2}\right) \cdot \frac{1}{1+\gamma}, \end{aligned} \tag{2.70}$$

其中$\chi(\boldsymbol{Y}) = \frac{1}{n_p} + \frac{1}{n_u} + \frac{1}{n_n}$。

下面基于定理 2.2 考察由算法 2 输出的学习器。记：

$$\mathcal{A} = \left\{f : f(x) = \sum_{i=1}^{T} \alpha^t \cdot h^t, \alpha^t \geq 0,\ h^t \in \mathcal{H}_{DS},\ T \in \mathbb{N}_+\right\}. \tag{2.71}$$

为模型权重非负条件下 [3]，所有可能由算法 2 产生的模型集合。通过以下定理将$co\left(\mathcal{H}_{DS}\right)$上的结论推广至假设空间$\mathcal{A}$。

定理 2.3. 给定训练数据集$\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$、$\rho > 0$及弱分类器的阈值备选集$\boldsymbol{T}_e$，设样例均由独立采样生成。对于算法 2 在$T$轮迭代后的任意输出函数$f(\boldsymbol{x}) \in \mathcal{A}$，以及任意$\delta \in (0,1)$，下式至少以$1-\delta$概率成立:

$$\begin{aligned} R_{0-1}^{PN}(f) \leq & \left(\exp\left(-\sum_{i=1}^{T} KL\left(\frac{1+\rho}{2} || \frac{1+\Delta^t}{2}\right)\right)\right. \\ & + \frac{8\sqrt{2}}{\rho}\left((\log d + \log K) \cdot \chi(\boldsymbol{Y})\right)^{1/2} \\ & \left. + \frac{1}{\rho}\left(\log(\frac{2}{\delta}) \cdot \chi(\boldsymbol{Y})\right)^{1/2}\right) \cdot \frac{1}{1+\gamma}, \end{aligned} \tag{2.72}$$

[3] $\alpha^t \geq 0$可通过设置较小的ρ或早停机制实现

另外对于任意满足$\min_t \Delta^t \geq \delta > \rho$的输出函数$f$，有：

$$\begin{aligned} R_{0-1}^{PN}(f) \leq & \left(\exp\left(-\frac{T}{2} \cdot (\delta - \rho)^2 \right) \right. \\ & + \frac{8\sqrt{2}}{\rho} \left((\log d + \log K) \cdot \chi(\boldsymbol{Y}) \right)^{1/2} \\ & \left. + \frac{1}{\rho} \left(\log(\frac{2}{\delta}) \cdot \chi(\boldsymbol{Y}) \right)^{1/2} \right) \cdot \frac{1}{1+\gamma}, \end{aligned} \tag{2.73}$$

证明. 给定$f = \sum_{i=1}^{T} \alpha^t \cdot h^t \in \mathcal{A}$，构造$\tilde{f} = \frac{f}{\sum_{i=1}^{T} \alpha^t}$，显然$\tilde{f} \in co(\mathcal{H}_{DS})$且$r_\rho(\tilde{f}) = r_\rho(f)$, $R_{0-1}^{PN}(f) = R_{0-1}^{PN}(\tilde{f})$，因此对$\mathcal{A}$的泛化界即可转化为对$\mathcal{F} = \left\{ \tilde{f} : f \in \mathcal{A} \right\}$的泛化界，又因为$\mathcal{F} \subseteq co(\mathcal{H}_{DS})$，因此可直接应用定理 2.2 结论。综合定理 2.1 及定理 2.2 结论，本定理即可得证。 □

注 2.1. 对于本定理，有以下结论:

1. 考察定理中两个不等式的基本形式，左端项为有监督情况下的泛化0-1误差，右端项与中弱分类器个数T有关的一项大致随T增大以指数形式衰减；右端项中与T无关项组成

$$\mathcal{O}\left(\left(\frac{1}{n_p} + \frac{1}{n_u} + \frac{1}{n_n} \right)^{1/2} \right)$$

的残差项，随训练样本个数增大而衰减。因此本定理即可证明式 (2.61)成立，也即当样本充分时，在半监督条件下运行算法 2 即可有效降低有监督情况下的泛化0-1误差。因此证明了本章方法的可行性。

2. 相比于目前已有的

$$\mathcal{O}\left((\frac{1}{n_p})^{1/2} + (\frac{1}{n_u})^{1/2} + (\frac{1}{n_n})^{1/2} \right)$$

半监督AUC泛化界(Sakai 等, 2018)，借助本章所提出的广义最大值不等式，即引理 A.6获得了量级更小的上界。

3. 本定理中的两个上界均为T的单减不增函数。因此本章算法的另一个优势在于：增加弱分类器个数在提升训练集性能的同时并不会带来明显的过拟合风险。

4. 定理中的两个不等式右端项与T无关部分反比于ρ，该性质表明引入ρ可降低泛化误差上界，为所提出算法引入正则项$\exp(\rho \cdot \sum_{i=1}^{T} \alpha^t)$提供了理论依据。

表 2.1 仿真数据集加速前后的运行时间对比

Table 2.1 The Running Time Comparison on Simulation Datasets Before and After Acceleration

实现方式	样本量						
	100	150	200	250	300	350	400
Before	0.2936	0.3818	0.4886	0.6100	0.7890	0.9600	1.1542
After	0.0227	0.0258	0.0278	0.0297	0.0356	0.0384	0.0405

表 2.2 数据集描述

Table 2.2 The description of datasets

数据集	数据来源	描述	样本数量	特征数量	正负样本比
pima	KEEL	糖尿病人分类数据	768	8	0.5360
ring	KEEL	多元正态分布数据集	7,400	20	0.9807
phoneme	KEEL	选手发音数据集	5,404	5	0.4154
vehicle0	KEEL	2D交通工具图像数据集	846	18	0.3075
vehicle2	KEEL	2D交通工具图像数据集	846	18	0.3471
credit-g	UCI	德国民众信用预测数据集	1,000	20	0.4285
glass1	UCI	玻璃杯种类预测数据集	218	9	0.5507
wisconsin	UCI	乳腺癌分类数据集	683	9	0.5382
wdbc	UCI	乳腺肿块分类数据集	569	30	0.5938
shuttle-c0-vs-c4	UCI	航班飞行数据集	58,000	9	0.0720
monk-2	UCI	MONK学习算法竞赛数据集	432	7	0.8947
sonar	UCI	声纳探测岩石数据集	208	60	0.8738
Surgical-deepnet	Kaggle	外科诊疗记录数据集	14,600	25	0.3371
insurance	Kaggle	健康护理保险数据集	381,000	11	0.1959
numerai	Kaggle	金融股票数据集	96,300	21	0.9753
cod-rna	Libsvm	RNA序列检测数据集	271,617	8	0.5000

2.5　实验

2.5.1　加速算法验证

为验证本文所提出加速方法的实际加速效果，生成不同规模的仿真数据集对加速前后的算法运行时间进行对比。具体而言，分别生成样本量为100,150,200,250,300,350,400的训练数据集。输入特征X每维按照正态分布$\mathcal{N}(0, 0.01)$, 维度为10。同时由分布$\mathcal{N}(0, 1)$ 逐维生成线性模型参数ω，总维度为10。进一步通过 $s = X\omega + \epsilon$生成得分函数，其中ϵ为服从$\mathcal{N}(0, 0.0001)$的白噪声。最终通过$y = \boldsymbol{I}[s > 0.1]$生成样本的标注结果。加速前后运行时间对比如表2.1。如表所示，本文加速算法可带来显著的计算效率提升。

2.5.2　数据集

表 2.3 测试集上AUC性能对比

Table 2.3 AUC performance on the test set[4]

数据集	Ours	RBAUC	PNUAUC	Samult	PNUAB	LSAUC
pima	**.8022**(**.0481**)	.7923(.0476)*	.6458(.0546)*	.6459(.0480)*	.6743(.0370)*	.6511(.0447)*
ring	**.9470**(**.0095**)	.9315(.0120)*	.9099(.0061)*	.9099(.0061)*	.8603(.0131)*	.9117(.0058)*
phoneme	**.8579**(**.0143**)	.8543(.0149)*	.8009(.0156)*	.8005(.0145)*	.7798(.0171)*	.7753(.0129)*
vehicle0	.9518(.0215)	.9465(.0207)	.9780(.0182)$^-$	**.9846**(**.0171**)$^-$	.8172(.0432)*	.8550(.0317)*
vehicle2	.9704(.0153)	.9595(.0228)*	.9647(.0195)	**.9810**(**.0134**)$^-$	.8555(.0417)*	.7588(.1247)*
credit-g	**.7290**(**.0314**)	.7267(.0420)	.6968(.0553)*	.7147(.0523)	.6296(.0597)*	.6785(.0813)*
glass1	**.7225**(**.0904**)	.7048(.0795)	.5828(.1199)*	.6365(.0898)*	.6321(.1086)*	.5833(.1304)*
wisconsin	**.9909**(**.0065**)	.9892(.0076)	.9195(.0282)*	.9211(.0275)*	.9105(.0257)*	.9054(.0384)*
wdbc	**.9873**(**.0100**)	.9859(.0100)	.9704(.0180)*	.9786(.0193)*	.8367(.0489)*	.9725(.0226)*
shuttle-c0-vs-c4	**1.000**(**.0000**)	.9801(.0322)*	.9998(.0005)	.9998(.0008)	.9526(.0354)*	.9922(.0176)*
monk-2	**.9844**(**.0202**)	.9839(.0196)	.8065(.0498)*	.8213(.0474)*	.9462(.0260)*	.7294(.0570)*
sonar	**.7247**(**.1128**)	.7140(.0768)	.6664(.1025)*	.6719(.1089)*	.5991(.0774)*	.5673(.0864)*
Surgical	**.8332**(**.0095**)	.8266(.0114)*	.7797(.0082)*	.7910(.0066)*	.7820(.0160)*	.6291(.0178)*
insurance	**.8810**(**.0012**)	.8765(.0015)*	.6395(.0072)*	.6457(.0065)*	.8290(.0022)*	.6081(.0198)*
numerai	**.5245**(**.0049**)	.5226(.0045)*	.5229(.0040)	.5229(.0040)	.5112(.0044)*	.5210(.0049)*
cod-rna	**.9681**(**.0019**)	.9362(.0018)*	.9313(.0018)*	.9363(.0017)*	.9218(.0020)*	.9314(.0018)*
总体均值	**.8672**	.8582	.7836	.8009	.8101	.7544
win/tie/lose	/	9/7/0	16/0/0	12/3/1	11/3/2	16/0/0

[4]其中对比方法标记$\star$表示本章所提出算法显著优于该算法，且其统计显著性经wilcoxson符号秩和检验p值小于0.05；对比方法标记$-$表示本章所提出算法性能劣于该算法，且其统计显著性经wilcoxson符号秩

为验证模型的有效性，本章在16个常见的二分类数据集上进行实验。数据集主要来源于keel[5]，UCI[6]，Kaggle[7]和Libsvm[8]。其中keel包含的数据集包括 vehicle2（2D交通工具图像数据集）、vehicle0（同左）、pima（糖尿病分类数据集）、ring（多元正态分布数据集）、phoneme（选手发音数据集）；UCI包含的数据集包括 credit-g（信用预测数据集）、glass1（玻璃分类数据集）、wisconsin（乳腺癌分类数据集）、wdbc（乳腺肿块分类数据集）、shuttle-c0-vs-c4（航班飞行数据集）、monk-2（MONK学习算法竞赛数据集）、sonar（声纳岩石探测数据集）；Kaggle包含的数据集包括 Surgical-deepnet（外科诊疗记录数据集）、insurance（健康护理保险数据集）和numerai （金融股票数据集）；Libsvm包含的数据集包括cod-rna（RNA序列检测数据集）。所有数据集的详细信息如表 2.2 所示。为验证本章算法在不同特性数据集上的表现，本章选用的数据集涵盖了较为宽泛的数据集规模（208-381,000）、不平衡比例(0.19-0.97)、以及应用场景（金融、保险、交通、医学、生物信息等）。同时，为测试模型在复杂数据集中的表现情况，本章在2016年于Kaggle发布的加密股市数据预测竞赛[9]使用的公开Numerai数据集上进行实验。Numerai数据集是基于对冲基金建立的股票数据集，该数据集采用了同态加密，所有的属性均被归一化，特征含义被隐藏，预测难度较大。

2.5.3 对比方法

为验证本章所提出算法的有效性，在实验中采用以下对比方法与本章算法进行对比：

- **RBAUC（Boosting集成算法+AUC优化）**：该方法采用RankBoost(Freund 等, 2003b)模型进行AUC优化，相比于本章提出算法，其训练过程仅利用训练集中所有已标记正负样本，而对未标记数据不加以利用。
- **PNUAB （半监督学习+Boosting集成算法）**：该方法基于AdaBoost(Freund 等, 1996)模型设计，为了适应半监督实验设定，本章将未标记数据进行复制，

和检验p值小于0.05；win/tie/loss 分别表示本章算法显著优于该算法且$p < 0.05$次数/本章算法与该算法差异无显著性次数/本章算法显著劣于该算法且$p < 0.05$次数；对于每个数据集上最优性能以**粗体**标出、次优性能以下划线标出。

[5] https://sci2s.ugr.es/keel/datasets.php

[6] http://archive.ics.uci.edu/ml/datasets

[7] https://www.kaggle.com

[8] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[9] https://www.kaggle.com/numerai/encrypted-stock-market-data-from-numerai

一份视作未标记的正样本，一份视作未标记的负样本。正样本、未标记正样本、负样本和未标记负样本的样本权重分别设为$\frac{1-\gamma}{2n_P}$、$\frac{\gamma}{2n_U}$、$\frac{\gamma}{2n_N}$和$\frac{1-\gamma}{2n_U}$，其中n_P、n_U和n_N分别为正样本、未标记样本和负样本的数量，$\gamma \in [0,1]$为实验中的超参数。相比与本章所提出的算法，相比于本章提出的算法，该算法没有直接利用AUC优化训练模型。

• **PNU-AUC(Sakai 等, 2018)（单模型半监督AUC优化算法）**：该方法同时利用正负样本以及未标记数据进行AUC优化，由于需要估计未标记样本中正负样本的比例，本章将未标记样本的真实正负样本比例作为输入，采用公开的代码[10]进行实验。相比于本章提出算法，该算法仅采用单个模型进行半监督AUC优化，未采用模型集成技术。

• **Samult(Xie 等, 2018a)（单模型半监督AUC优化算法）**：该对比方法同时利用正负样本以及未标记数据进行AUC优化而无需估计未标记样本中正负样本的比例，本章对Samult模型进行复现从而进行实验。相比于本章提出算法，该算法仅采用单个模型进行半监督AUC优化，未采用模型集成技术。

• **LSAUC （基线算法）**：LSAUC采用平方损失作为替代损失，替代优化问题的解由闭式解直接得出，同时该算法仅利用训练集中所有已标记正负样本，对未标记数据不加以利用。由于LSAUC算法既未利用模型集成技术也未利用半监督学习技术，因此为本实验的基线算法。

2.5.4　实验细节

本章采用分层采样，等比地将正负样本中的70%，15%和15%的样本分别划分为训练集，验证集和测试集。同时，为了符合半监督实验设定，随机将训练集中85%的数据的标签去掉，作为未标记数据。最后，为了提高实验的可靠性，独立进行了15次训练集、验证集、测试集的采集，根据模型在验证集上的15次结果的均值来进行模型选择，并将对应的测试集上的15次结果的均值作为最后评估模型的基准。

2.5.5　实验结果

各算法在16个数据集上的实验结果如表2.3所示。从表 2.3 可以看出，本章所提方法在绝大多数情况下都以0.05显著性水平优于其他方法。在全部16个数

[10] https://github.com/t-sakai-kure/pywsl

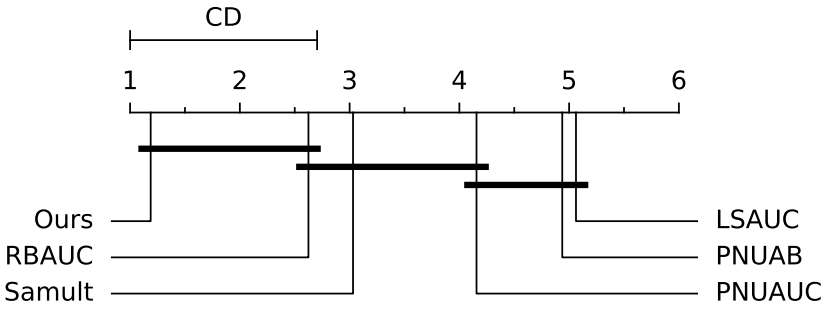


图 2.1 各算法差异对比图

Figure 2.1 The comparison between different methods

据集中，本章所提方法在14个数据集上都表现最优。从各数据集上的平均性能来看，本章所提方法也达到了最好效果。此外，给出更加详细的对比分析。

所提方法 vs. 半监督+AUC优化： PNU-AUC和Samult都是半监督学习和AUC优化结合所产生的方法。所提方法和PNU-AUC相比，在各个数据集上平均性能提升8.31%。和Samult相比，平均性能提升7.94%。其中，在pima数据集上相较于PNU-AUC提升24.22%，比Samult提升24.20%；在ring数据集上比PNU-AUC提升4.08%，比Samult提升4.08%；在phoneme数据集上比PNU-AUC和Samult分别提升7.12%和7.17%；在vehicle0数据集上相较于这两种方法分别降低2.68%和3.33%；在vehicle2数据集上比PNU-AUC提升0.59%，比Samult降低1.08%；在credit-g数据集上比PNU-AUC提升4.62%，比Samult提升2.00%；在glass1数据集上相较于PNU-AUC提升23.97%，比 Samult提升13.51%；在wisconsin数据集上比这两种方法分别提升7.77%和提升7.58%；在wdbc数据集上比PNU-AUC提升1.74%，比Samult提升0.89%；在shuttle-c0-vs-c4数据集上比两种方法都提升了0.02%；在 monk-2数据集上比PNU-AUC提升22.06%，比Samult提升19.86%；在sonar数据集上比两种方法分别提升8.75%和7.86%；在Surgical数据集上比两种方法分别提升6.86%和 5.34%；在insurance数据集上比PNU-AUC提升37.76%，比Samult提升36.44%；在numerai数据集上比PNU-AUC提升0.31%，比Samult提升0.31%；在cod-rna数据集上比PNU-AUC提升3.95%，比Samult提升3.40%。相比PNU-AUC，本章算法有12次性能提升是在0.05显著性水平之上。相比Samult，本章算法有11次性能提升的显著性水平超过0.05。上述实验结果有效验证了定理2.2中的结论，即，将boosting方法引入半监督AUC优化可在保证泛化能力的同时提升优

化收敛速率，进而带来提升总体性能提升。

所提方法 vs. Boosting+AUC优化：RBAUC同时利用了boosting方法及AUC优化方法。和该方法相比，本章所提方法带来的平均性能提升为 1.44%。在各个数据集上的提升分别为：在pima数据集上提升1.25%；在ring数据集上提升1.66%；在phoneme数据集上提升0.42%；在vehicle0数据集上提升0.56%；在vehicle2数据集上提升1.14%；在credit-g数据集上提升0.32%；在glass1数据集上提升2.51%；在wisconsin数据集上提升0.17%；在wdbc数据集上提升0.14%；在shuttle-c0-vs-c4数据集上提升2.03%；在hepatitis数据集上提升10.72%；在monk-2数据集上提升0.05%；在sonar数据集上提升1.50%；在Surgical数据集上提升0.80%；在 insurance数据集上提升0.51%；在numerai数据集上提升0.36%；在cod-rna数据集上提升3.41%。其中，有9次性能提升是在0.05显著性水平之上。实验结果表明在boosting算法中引入半监督学习能够有效利用未标注数据中的额外信息，从而提升模型在少量已标注数据上的学习效果。

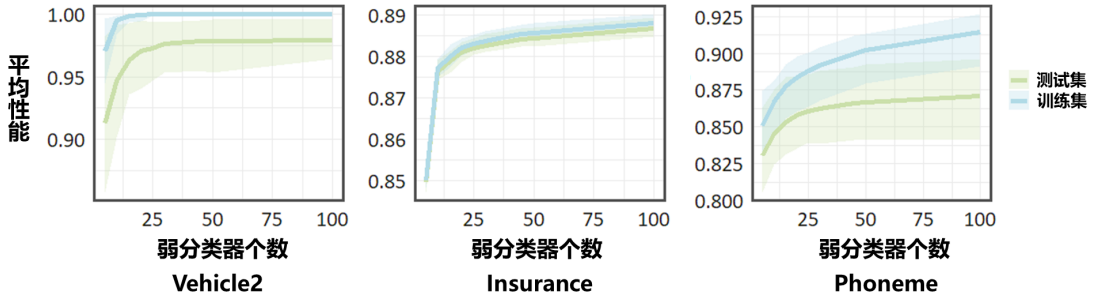


图 2.2 算法性能随弱分类器个数增加的变化趋势

Figure 2.2 The performan of algorithms *w.r.t.* the number of weak classifiers

PNUAB vs. 其它半监督方法： PNUAB是半监督学习和AdaBoost的结合。与其它半监督方法的平均进行对比， PNUAB在各数据集上的平均性能更低。以本章所提方法为例，平均性能比PNUAB高出10.79%，其中，有16次性能提升是在0.05显著性水平之上。该实验观测表明，在半监督条件下下进行AUC优化可有效地提高AUC泛化性能。

LSAUC vs. 其它方法： 其它方法和LSAUC相比，性能提升均较为明显。以本章所提方法为例，平均性能提升14.76%。各个数据集上的提升分别为：在pima数据集上提升23.21%；在ring数据集上提升3.87%；在phoneme数据集上提升10.65%；在vehicle0数据集上提升11.32%；在vehicle2数据集上提升27.89%；

在credit-g数据集上提升7.44%；在glass1数据集上提升23.86%；在wisconsin数据集上提升9.44%；在wdbc数据集上提升1.52%；在shuttle-c0-vs-c4数据集上提升了 0.79%；在hepatitis数据集上提升10.72%；在monk-2数据集上提升34.96%；sonar 数据集上提升27.75%；在Surgical数据集上提升32.44%；在insurance数据集上提升44.88%；在numerai数据集上提升0.67%；在cod-rna数据集上提升3.94%。其中有16次性能提升是在0.05显著性水平之上。实验结果表明半监督学习以及boosting 策略都均能有效克服由未标注数据所带来的性能瓶颈。

算法总体比较：为比较各个算法总体差异的统计显著性，对各个算法在各个数据集上的性能排序进行了事后假设检验，结果如图2.1,其中刻度中数字代表算法排序，水平线代表CD距离（crtical difference）。仅当两算法排序差距超过CD时，其差异具有显著性。如图所示，本章所提算法获得了最高排序，其他算法在图中距本章算法距离超过CD（crtical difference）长度即可视为具有显著性差异（$p < 0.05$）。可见除RBAUC之外，其余算法均与本章算法存在显著差异。另一方面，RBAUC与本章算法的差距亦接近CD距离。由此可见本章所提算法在整体水平也具有显著性优势。

弱分类器个性的影响：以vehicle2、insurance、phoneme三个数据集为例，图2.2 中验证了本章算法训练集和测试集平均AUC性能（越大越佳）随弱分类器增加的变化趋势，其中曲线阴影部分对应15次重复试验上的极差波动。如图所示，随着弱分类器的增加训练集的性能快速提升，并趋于收敛；同时，测试集上性能变化趋势与训练集基本一致，未出现过拟合现象。尤其对于vehicle2数据集，当弱分类器个数达到25左右时，其训练集性能很快到达最佳值1.00，而测试集在其后的训练过程中未出现下降趋势。综上分析可推知，本章所提出算法可兼顾模型的收敛速度以及泛化能力，再次验证了本章定理 2.3 的有效性。

2.6 小结

本章对半监督条件下基于Boosting 算法的AUC优化模型集成进行了系统性研究。在算法层面，提出一种高效的基于Boosting的AUC优化模型集成方法，可将单次迭代的空间/时间复杂度由$O(n_p n_n + n_p n_u + n_u n_n)$降至$O(n_p + n_u + n_n)$。进一步的理论分析证明，本章所提出算法可使训练集误差随弱分类器的增加以几何速度衰减且泛化误差随训练样本增加逐渐趋于0。在泛化误差分析方面，本章

通过构造半监督AUC的Rademacher复杂度以及广义最大值不等式给出了相比与文献(Sakai 等, 2018)更为紧致的上界。实验分析方面，本章在16个数据集上对所提方法的性能进行了验证。实验结果证实本章所提出算法可显著提升半监督学习条件下的模型AUC性能。

第3章　基于M度量的多分类AUC优化理论及方法

3.1　引言

在过去二十年中，AUC优化方法主要局限于二分类场景，如(Alan 等, 2004; Joachims, 2005, 2006; Calders 等, 2007; Narasimhan 等, 2013a; Gao 等, 2013; Narasimhan 等, 2017)，由于真实场景下的模式识别问题往往涉及两个以上类别，因此自然引出如何将AUC泛化至多类别场景的问题。针对该问题，本章探索多类场景下AUC诱导的机器学习框架。具体而言，提出一种通用的经验替代风险最小化框架，给出一致性、泛化能力的理论分析并构造高效损失、梯度加速算法，提升训练效率。

首先，本章对(Hand 等, 2001)提出的基于M度量的AUC多分类拓展进行分析。具体地，分析M度量如何有效地避免类排序对之间的不平衡问题，证明M度量作为多分类AUC指标的合理性。基于此，本章提出最小化由M度量（记为MAUC$^{\downarrow}$）诱导的0-1错排损失。基于所构造的错排损失，明确实现过程的三点主要挑战：**(a)**0-1错排损失MAUC$^{\downarrow}$离散、不可微，难以进行有效的优化；**(b)**数据先验分布未知，难以计算其期望值；**(c)**损失和梯度函数的复杂度极高，无法进行端到端训练。

针对**(a)**，本章研究如何构造0-1风险的替代风险。为此，推导 MAUC 准则下的贝叶斯最优得分函数集合，并且证明主流的替代损失函数在特定假设下是Fisher一致的，即优化相应的替代期望风险也可得到贝叶斯最优得分函数。

针对**(b)**，本章构建一个MAUC准则的经验替代风险最小化框架。将训练数集上替代风险的平均经验估计作为目标函数，以规避对期望损失的计算。此外，对该方法的泛化能力进行系统地分析。其中主要挑战在于：经验风险函数不能分解为独立损失函数的有限和，导致传统的对称化技术不可用(Mohri 等, 2018)。为解决该问题，本章从MAUC$^{\downarrow}$自身特点出发，构造其Rademacher复杂度，并通过该复杂度给出常见深度模型的泛化误差上界。该理论结果对于数据不平衡性具有感知能力，比传统结果更关注造成性能瓶颈的少数类样本。

针对**(c)**，构造加速算法以提升可拓展性，使算法能够在小批量/全批量场景下以线性复杂度计算损失和梯度。

最后，在11个真实数据集上进行实验，以验证所框架的有效性。

3.2 预备基础

3.2.1 符号定义

3.2.1.1 基本符号

本章主要关注两类随机事件：对于成对AUC指标，$\mathcal{E}^{(ij)}$表示事件$y_1 = i, y_2 = j \vee y_1 = j, y_2 = i$，而$\mathcal{E}^{(i)}$表示事件 $y_1 = i, y_2 \neq i \vee y_1 \neq i, y_2 = i$。给定事件$\mathcal{A}$，$I[\mathcal{A}]$ 是该事件的指示函数：当$\mathcal{A}$成立时函数值为1，反之则为0。给定有限集合$\mathcal{S}$，N_C表示类别数，N表示数据集中的总样本数。

3.2.1.2 基本设定

对于类别数为N_C的多类问题，所有样本均采样自乘积空间 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$；$\mathcal{X}$为输入特征$\mathcal{X} \subset \mathbb{R}^d$； d为输入特征维度；$\mathcal{Y}$为标签空间$[N_C]$。对于任意给定标签$y_m = i$，将其表示为one-hot向量 $\boldsymbol{y}_m = [y_m^{(1)}, \cdots, y_m^{(N_C)}]$，仅当$k \neq i$时$y_m^{(k)} = 0$，否则$y_m^{(k)} = 1$。本章采用 one *vs.* all 多类别分解策略(Mohri 等, 2018)。更为具体地，将N_C个类别的（$N_C > 2$）得分函数表示为N_C个函数 $f = (f^{(1)}, \cdots, f^{(N_C)})$，其中 $f^{(i)} : \mathcal{X} \to \mathbb{R}$ 为表示$y = i$置信度的连续得分函数。

3.2.1.3 二分类AUC

在二类问题中，AUC具有明确的统计意义: 其等价于Wilcoxon 统计量(Hanley 等, 1982)，即任意一对正负样本具有正确偏序关系的概率：

$$\begin{aligned}\mathrm{AUC}(f) &= \mathbb{P}\left[\Delta(y)\Delta(f) > 0 | \mathcal{E}^{(0,1)}\right] + \frac{1}{2}\mathbb{P}\left[\Delta(f) = 0 | \mathcal{E}^{(0,1)}\right] \\ &= \mathop{\mathbb{E}}_{z_1 \sim \mathcal{D}_{\mathcal{Z}}, z_2 \sim \mathcal{D}_{\mathcal{Z}}} \left[\boldsymbol{I}[\Delta(y)\Delta(f) > 0] + \frac{1}{2}\boldsymbol{I}[\Delta(f) = 0] | \mathcal{E}^{(0,1)}\right],\end{aligned}$$

其中$\Delta(y) = y_1 - y_2$， $\Delta(f) = f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)$。注意，本章依据惯例(Clémençon 等, 2008; Agarwal, 2014; Gao 等, 2015)将正负例持平的得分定义为0.5。

3.2.2 研究动机

首先，需要确定合适的多类度量。如第3.2.1节所述，考虑以下两种备择分解方式将多类AUC构造为多个二类AUC的平均值。

3.2.2.1 One *vs.* All (ova) 转换

给定一个ova得分函数$f = (f^{(1)}, \cdots, f^{(N_C)})$， 可为每个$f^{(i)}$构建一个对应

的AUC损失，其中正例为当前类样本，负例为其他类样本，并通过类平均AUC表示最终的AUC损失：

$$\mathrm{AUC}^{\mathrm{ova}}(f)=\frac{1}{N_C}\sum_{i=1}^{N_C}\mathrm{AUC}_{i|\neg i}(f^{(i)}), \tag{3.1}$$

其中：

$$\mathrm{AUC}_{i|\neg i}(f^{(i)})=\mathop{\mathbb{E}}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\boldsymbol{I}[\Delta(y_{1,2}^{(i)})\Delta(f^{(i)})>0|\mathcal{E}^{(i)}]+\frac{1}{2}\boldsymbol{I}[\Delta(f^{(i)})=0]|\mathcal{E}^{(i)}\right]$$
$$\Delta(y_{1,2}^{(i)})=y_1^{(i)}-y_2^{(i)},\ \Delta(f^{(i)})=f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)$$

注意，本章在此处采取等权重平均值以避免不平衡问题。

3.2.2.2　One *vs*. One (ovo) 转换（M度量）

根据(Hand 等, 2001)，可将多类AUC指标表示为每个类别对(i,j)的二类AUC分数的平均值，其定义为：

$$\mathrm{AUC}^{\mathrm{ovo}}(f)=\frac{\sum_{i=1}^{N_C}\sum_{j\neq i}\mathrm{AUC}_{i|j}(f^{(i)})}{N_C(N_C-1)}, \tag{3.2}$$

其中，$\mathrm{AUC}_{i|j}(f^{(i)})=\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\boldsymbol{I}[\Delta(y_{1,2}^{(i)})\Delta(f^{(i)})>0|\mathcal{E}^{(ij)}]+\frac{1}{2}\boldsymbol{I}[\Delta(f^{(i)})=0]|\mathcal{E}^{(ij)}\right]$。此处需注意$\mathrm{AUC}_{i|j}\neq\mathrm{AUC}_{j|i}$，因二者使用不同的得分函数。

3.2.2.3　比较性质

为在上述两种分解策略中选出较优者，给出以下定理对$\mathrm{AUC}^{\mathrm{ovo}}$及$\mathrm{AUC}^{\mathrm{ova}}$进行比较。

定理 3.1 (**对比性质**). 给定标签分布$\mathbb{P}[y=i]=p_i>0$以及多类得分函数f，以下性质成立:

(a)

$$\mathrm{AUC}^{\mathrm{ova}}(f)=\frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{j\neq i}\left(\frac{p_j}{1-p_i}\right)\cdot\mathrm{AUC}_{i|j}(f^{(i)}) \tag{3.3}$$

(b)

$$\mathrm{AUC}^{\mathrm{ova}}(f)=\mathrm{AUC}^{\mathrm{ovo}}(f),\ \text{当}\ p_i=\frac{1}{N_C},\quad i=1,2,\cdots,N_C \tag{3.4}$$

(c) 当且仅当$\mathrm{AUC}^{\mathrm{ovo}}(f)=1$时$\mathrm{AUC}^{\mathrm{ova}}(f)=1$。

定理3.1-(a)表明$\mathrm{AUC}^{\mathrm{ova}}$将$\mathrm{AUC}_{i|j}$赋予权重$\frac{p_j}{1-p_i}$，数值正比于类对$(i,j)$的样本比重，导致忽略少数类别对的性能，不利于不平衡数据集的学习。相反，$\mathrm{AUC}^{\mathrm{ovo}}$则对所有成对AUC赋予相等的权重，近而更好的缓解类频对性能的影响。定理3.1-(b)表明当标签分布接近平衡时，$\mathrm{AUC}^{\mathrm{ova}}$和$\mathrm{AUC}^{\mathrm{ovo}}$趋于相等。定理3.1-(c)进一步证明当$f$使性能最大化时$\mathrm{AUC}^{\mathrm{ova}}$和$\mathrm{AUC}^{\mathrm{ovo}}$趋于一致。由此可见，$\mathrm{AUC}^{\mathrm{ovo}}$相比$\mathrm{AUC}^{\mathrm{ova}}$更适合类不平衡的长尾数据集。

3.2.3 研究目标

本章旨在构造最大化$\mathrm{AUC}^{\mathrm{ovo}}$的学习算法。为适应标准机器学习范式，本章依照惯例(Clémençon 等, 2008; Gao 等, 2015; Agarwal, 2014; Narasimhan 等, 2013b,c)，将最大化问题转化为基于期望风险的最小化问题 $f \in \operatorname{argmin}_f R(f)$，

$$R(f) = \mathrm{MAUC}^{\downarrow} = \sum_{i=1}^{N_C} \sum_{j \neq i} \frac{\mathbb{E}_{z_1,z_2}[\ell_{0-1}^{i,j,1,2} | \mathcal{E}^{(ij)}]}{N_C(N_C-1)}, \tag{3.5}$$

其中，

$$\begin{aligned} \ell_{0-1}^{i,j,1,2} &= \ell_{0-1}(f^{(i)}, \boldsymbol{x}_1, \boldsymbol{x}_2, y_1^{(i)}, y_2^{(i)}), \\ &= \boldsymbol{I}\left[\Delta(y_{1,2}^{(i)}) \cdot \Delta(f^{(i)}) < 0\right] + \frac{1}{2}\boldsymbol{I}\left[\Delta(f^{(i)}) = 0\right] \end{aligned} \tag{3.6}$$

为0-1错排损失。

3.3 一致性分析

由于0-1错排损失不可微，直接求解优化问题极为困难。本节将讨论如何构造替代风险$R_\ell(f)$，并以替代损失ℓ作为0-1错排损失的可微替代函数。

首先推导出需要近似的目标：理想条件下，最下化0-1损失时对应的最优函数类。换言之，需得出满足

$$f \in \operatorname*{argmin}_f R(f)$$

的所有函数的基本形式。根据机器学习领域的惯例，将此类函数称为**贝叶斯最优得分函数**。严格意义上，贝叶斯最优得分函数应为所有可测函数中$R(f)$最小的函数集合。但由于本章主要专注于分类问题，需对得分函数进行适当归一化，因此本章将最小化限定于以下函数类内：

$$f \in \mathcal{F}_\sigma^{N_C} = \underbrace{\mathcal{F}_\sigma \times \mathcal{F}_\sigma \cdots \times \mathcal{F}_\sigma}_{N_C},$$

其中 $\times$ 为集合的笛卡尔积，$\mathcal{F}_\sigma$由下式给出：

$$\mathcal{F}_\sigma = \{g : g \text{ 为值域为 } [0,1] \text{ 可测函数}\} \tag{3.7}$$

其中得分函数输出限制在$[0,1]$内。在此限定基础上，本章记**贝叶斯最优得分函数**为所有满足以下条件的得分函数：

$$f \in \underset{f \in \mathcal{F}_\sigma^{N_C}}{\operatorname{argmin}} R(f),$$

在此基础上，由以下定理给出数据先验分布已知情况下的贝叶斯最优得分函数数学形式。

定理 3.2 (**贝叶斯最优得分函数**). 给定$\eta_i(\cdot) = \mathbb{P}[y = i|x]$和$p_i = \mathbb{P}[y = i]$，可得出以下结论:

(a) 若

$$\Delta(f^{(i)}) \cdot \Delta(\pi) > 0,\ \forall \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \neq \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1), \tag{3.8}$$

其中，

$$\begin{aligned}
&\Delta(f^{(i)}) = f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2) \\
&\Delta(\pi) = \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) - \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1) \\
&\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j \neq i} \frac{\eta_i(\boldsymbol{x}_1)\eta_j(\boldsymbol{x}_2)}{2p_i p_j}, \\
&\pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1) = \sum_{j \neq i} \frac{\eta_j(\boldsymbol{x}_1)\eta_i(\boldsymbol{x}_2)}{2p_i p_j},
\end{aligned} \tag{3.9}$$

则$f = \{f^{(i)}\}_{i=1,2,\cdots,N_C}$是MAUC$^{\downarrow}$准则下的贝叶斯最优得分函数。

(b) 令$\sigma(\cdot)$表示sigmoid函数，$s_i(\boldsymbol{x}) = \eta_i(\boldsymbol{x})/p_i$，$s_{\setminus i}(\boldsymbol{x}) = \sum_{j \neq i} s_j(\boldsymbol{x})$，则贝叶斯最优得分函数可表示为:

$$f^{\star(i)}(\boldsymbol{x}) = \begin{cases} \sigma\left(\dfrac{s_i(\boldsymbol{x})}{s_{\setminus i}(\boldsymbol{x})}\right), & 0 \leq \eta_i(\boldsymbol{x}) < 1 \\ 1, & \eta_i(\boldsymbol{x}) = 1. \end{cases} \tag{3.10}$$

该定理除可给出贝叶斯最优得分函数的形式之外，还表明最优得分函数可按照 $\left(\frac{s_i(\boldsymbol{x})}{s_{\setminus i}(\boldsymbol{x})}\right)$对$f^{(i)}(\boldsymbol{x})$样本进行排序。该变量可被视为$y = i$与$y \neq i$的置信度比值。与传统方法中直接采用后验概率$\mathbb{P}(y = i|\boldsymbol{x})$作为置信度的方式不同，此处后验分布 $\mathbb{P}(y = i|\boldsymbol{x})$由因子$1/p_i$加权，且$\frac{\mathbb{P}(y=i|\boldsymbol{x})}{p_i}$ 正比于类频无关概率：$\mathbb{P}(\boldsymbol{x}|y = i)$，因此MAUC$^{\downarrow}$意义下的贝叶斯最优得分函数不受类频影响，对于长尾数据具有一定优势。

由于数据分布不可知，以及0-1损失不可微，即使采用定理3.2 中的解决方案仍然难以处理贝叶斯得分函数。因此，需要将0-1损失替换为凸可微损失函数ℓ，并近似求解以下问题：

$$f^{\star} \in \operatorname*{argmin}_{f} R_{\ell}(f),$$

$R_{\ell}(f)$ 为替代形式的期望风险:

$$R_{\ell}(f) = \sum_{i} \frac{R_{\ell}^{(i)}(f^{(i)})}{N_C(N_C - 1)}, \tag{3.11}$$

$R_{\ell}^{(i)}$为第i个得分函数对应的期望风险：

$$R_{\ell}(f^{(i)}) = \sum_{j \neq i} \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2} \left[\ell(\Delta(y_{1,2}^{(i)}) \Delta f^{(i)}) | \mathcal{E}^{(ij)} \right]. \tag{3.12}$$

为确保近似求解的有效性，替代损失需要满足与MAUC的渐进一致性。换言之，应确保通过最小化该替代风险可在渐进意义下得到贝叶斯最优函数。

由此引出极限条件下的一致性定义。

定义 3.1 (MAUC$^{\downarrow}$**一致性**). [1] 如果对于任意函数序列$\{f_t\}_{k=1,2,\ldots}$均有：

$$R_{\ell}(f_t) \to \inf_{f \in \mathcal{F}_{\sigma}^{N_C}} R_{\ell}(f) \text{ 蕴含 } R(f_t) \to \inf_{f \in \mathcal{F}_{\sigma}^{N_C}} R(f). \tag{3.13}$$

那么称ℓ与MAUC$^{\downarrow}$一致。

基于该定义，进一步给出一致性的充分条件，结论如下定理，详细证明见附录B.2.2。

定理 3.3 (MAUC$^{\downarrow}$ **一致性**). 若替代损失函数ℓ在区间$[-1, 1]$为可导非增凸函数，且$\ell'(0) < 0$，则替代损失 ℓ 与 MAUC$^{\downarrow}$ 一致。

[1]若无特别说明，所有分布$\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$都满足此条件。

推论 3.1. 由定理3.3，以下结论成立:

1. Logit 损失 $\ell_{logit}(x) = \log(1+\exp(-x))$ 与 MAUC$^{\downarrow}$ 一致。
2. 指数损失 $\ell_{\exp}(x) = \exp(-x)$ 与 MAUC$^{\downarrow}$一致。
3. 平方损失 $\ell_{sq}(x) = (1-x)^2$ 与 MAUC$^{\downarrow}$一致。
4. q-范数($q > 1$)铰链损失 $\ell_q(x) = (\max(1-x, 0))^q$ 是 MAUC$^{\downarrow}$一致的。
5. 对任意$1/2 > \epsilon > 0$，广义铰链损失:

$$\ell_{\epsilon,m}(x) = \begin{cases} m - t, & t \le 1-\epsilon \\ (t-1-\epsilon)^2/4\epsilon, & 1-\epsilon \le t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

与 MAUC$^{\downarrow}$一致。

6. 对任意$1 > \epsilon > 0$距离加权损失:

$$\ell_d(x) = \begin{cases} 1/t, & t > \epsilon \\ 1/\epsilon \cdot (2 - t/\epsilon), & \text{otherwise} \end{cases} \tag{3.15}$$

与 MAUC$^{\downarrow}$ 一致。

由于铰链损失 $\ell_{hinge}(t) = \max(1-t, 0) = \lim_{\epsilon \to 0} \ell_\epsilon(t)$，因此铰链损失至少为贝叶斯一致损失函数类的极限点, 近似具有一致性。

3.4　经验风险最小化

本节进一步针对无法估计期望风险的问题，构建经验风险最小化框架，并对其泛化性能进行分析。

3.4.1　经验替代风险最小化

为避免直接计算期望风险$R_\ell(f)$，通常可从数据分布中经过有限次采样得到训练数据$\mathcal{S} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$，并进行风险估计。为实现这一目的，以下命题给出$R_\ell(f)$在样本$\mathcal{S}$的无偏估计。其中，为表示$\mathcal{S}$中的标签频率，以 $\mathcal{N}_i = \{\boldsymbol{x}_k : y_k = i, z_k \in \mathcal{S}\}$表示具有标签$i$的样本集。定义$n_i = |\mathcal{N}_i|$为属于$\mathcal{S}$中第$i$类的实例数。

命题 3.1 (**无偏估计**). 定义$\hat{R}_{\ell,\mathcal{S}}(f)$为:

$$\hat{R}_{\ell,\mathcal{S}}(f) = \sum_{i=1}^{N_C} \sum_{j \ne i} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \ell^{i,j,m,n},$$

$\ell^{i,j,m,n}$是$\ell(f^{(i)}(\boldsymbol{x}_m)-f^{(i)}(\boldsymbol{x}_n))$的简写，则$\hat{R}_{\ell,\mathcal{S}}(f)$为$R_\ell(f)$的无偏估计，满足$R_\ell(f)=\mathop{\mathbb{E}}_{\mathcal{S}}(\hat{R}_{\ell,\mathcal{S}}(f))$。

根据上述命题结论，可通过最小化有限训练集$\mathcal{S}$上的经验风险$\hat{R}_{\ell,\mathcal{S}}(f)$近似完成风险最小化。在实际问题中，得分函数$f$通常可通过参数$\theta$刻画，因此将$f$记作$f_\theta$（如，线性模型可定义为$f_\theta(\boldsymbol{x})=\theta^\top\boldsymbol{x}$）。此外，为防止过拟合，通常将$f$的选择限制在特定假设类$\mathcal{H}$内。同样地，$\mathcal{H}$本质上是参数化函数$f_\theta$的集合，因此限制$f$于$\mathcal{H}$内，等价于限制$\theta$限制在某参数集合$\Theta$内，其中$\mathcal{H}$及$\Theta$存在特定的对应关系（如，$\mathcal{H}$可以定义为满足$||\theta|| \leq \gamma$的所有线性模型）。利用该假设类，可以构造以下优化问题，在$\mathcal{H}$之中学习f：

$$\min_{f\in\mathcal{H}} \hat{R}_{\ell,\mathcal{S}}(f). \tag{3.16}$$

从参数角度来看，式 (3.16)等价于求解令目标最小化的参数$\theta\in\Theta$。此外，参数约束$\theta\in\Theta$也可近似表示为正则项 $Reg_\Theta(\theta)$。因此，可通过如下问题近似优化式(3.16):

$$\min_{\theta} \hat{R}_{\ell,\mathcal{S}}(f_\theta)+\alpha\cdot Reg_\Theta(\theta). \tag{3.17}$$

式 (3.17)已将原问题充分简化，可通过机器学习技术进行求解。为方便数学表达，将采用符号f和$\mathcal{H}$，而不再显式定义θ。

3.4.2 MAUC$^{\downarrow}$ Rademacher复杂度及其性质

进一步地，研究此近似方法的泛化性能。一方面，一致的替代损失ℓ可以确保最小化替代风险即可找到贝叶斯最优得分函数。另一方面，在经验风险较小的情况下，还需要保证训练集性能可以较好地泛化到未知样本，即促使$R_\ell(f)$期望值足够小。为确保第二点，给定模型备选假设空间$\mathcal{H}$，本节将基于给出最坏情况下的泛化性能分析。具体而言，需确保不等式$R_\ell(f)\leq\hat{R}_\ell(f)+\delta$以大概率成立，且当$N$趋向无穷大时，$\delta$趋于零。

模型泛化能力通常依赖于所选定假设集$\mathcal{H}$的复杂程度，$\mathcal{H}$越为复杂往往模型的训练能力越强，但过拟合风险也随之增高。因此为完成本章方法的泛化性能分析，首先给出MAUC$^{\downarrow}$经验风险最小化框架下$\mathcal{H}$的复杂度程度刻画。本章中，采用机器学习理论中广泛使用的Rademacher复杂度(Mohri 等, 2018)作为复杂程度的度量。

不同于只涉及独立样本层次损失的传统机器学习问题，AUC的成对形式使得Rademacher复杂度难以刻画。具体地，MAUC公式中的被求和项间并不相互独立，[1]，导致标准的对称化技术(Mohri 等, 2018; Bartlett 等, 2002)不适用于此问题。为解决该问题，本章为MAUC$^{\downarrow}$损失给出Rademacher复杂度的扩展形式，定义如下。

定义 3.2 (MAUC$^{\downarrow}$ **Rademacher复杂度**). 给定数据集$\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$，经验MAUC$^{\downarrow}$ Rademacher 复杂度，以及假设空间$\mathcal{H}$ 定义如下：

$$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})=\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}T^{i,j,m,n}\right], \tag{3.18}$$

$$T^{i,j,m,n}=\frac{\sigma_m^{(i)}+\sigma_n^{(j)}}{2}\cdot\frac{\ell(f^{(i)}(\boldsymbol{x}_m)-f^{(i)}(\boldsymbol{x}_n))}{n_in_j},$$

对于$i=1,2,\cdots,N_C$，$\sigma_1^{(i)},\cdots,\sigma_{n_i}^{(i)}$是独立同分布的Rademacher随机变量。基于数量版本的MAUC$^{\downarrow}$ Rademacher复杂度可定义为

$$\mathfrak{R}_{\mathrm{MAUC}^{\downarrow}}(\ell\circ\mathcal{H})=\mathbb{E}_{\mathcal{S}}\left[\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})\right]$$

在此基础上由Rademacher复杂度诱导出泛化性能上界，如以下定理所示。证明见附录B.5。

定理 3.4 (**泛化界一般形式**). 给定数据集$\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$，样本由独立采样得到，对于所有多类得分函数$f\in\mathcal{H}$，若替代损失函数$\ell$的值域包含于$[0,B]$,，$\forall\delta\in(0,1)$，以下不等式至少依概率$1-\delta$成立：

$$R_\ell(f)\leq\hat{R}_{\mathcal{S}}(f)+C_1\cdot\frac{\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})}{N_C(N_C-1)}+C_2\cdot\frac{B}{N_C}\cdot\xi(\boldsymbol{Y})\cdot\sqrt{\frac{\log(\frac{2}{\delta})}{N}}, \tag{3.19}$$

C_1,C_2为常数，$\xi(\boldsymbol{Y})=\sqrt{\sum_{i=1}^{N_C}\frac{1}{\rho_i}}$,$\rho_i=\frac{n_i}{N}$。

定理中ρ_i表示第i类样本所占比例，而$\xi(\boldsymbol{Y})$刻画了样本不平衡程度，仅当ρ_i呈现均匀分布时，$\xi(\boldsymbol{Y})$达到最大化。因此该定理说明，误差期望$R_\ell(f)$随训练误差$\hat{R}_{\mathcal{S}}(f)$增大，随类别数N_C、样本量N、模型不平衡程度减小而收敛至0。

根据定理B.5.2，只需计算出经验Rademacher复杂度$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})$的适当上界，即可得到具体假设空间下的泛化界。为进一步简化构造Rademacher复

[1]如，当$\boldsymbol{x}_1=\boldsymbol{x}_1'$或$\boldsymbol{x}_2=\boldsymbol{x}_2'$时，$\ell(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2))$ 和$\ell(f^{(i)}(\boldsymbol{x}_1')-f^{(i)}(\boldsymbol{x}_2'))$ 相互依赖。

杂度的难度，以下将讨论基于覆盖数和chaining技术的$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})$上界构造通用技术。借助以下各定理，可将$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})$的上界转换为更简单模型类的Rademacher复杂度上界，避免对成对损失函数的处理。在本章后续内容中，定理B.5.4-(a)用于证明定理3.6；定理3.6用于证明定理3.7；定理B.5.4-(b)则用于定理3.8的证明之中。

基于附录B.5中的引理B.8所证明次高斯性质，可通过覆盖数推导出$\mathrm{MAUC}^{\downarrow}$ Rademacher复杂度上界，其定义可见定义3.3和定义3.4。

定义 3.3 (ϵ-覆盖). (Ledoux 等, 2013) 令$(\mathcal{H},d)$为一个(伪)度量空间，并且$\Theta\in\mathcal{H}$。当$\Theta\in\bigcap_{i=1}^{K}\mathcal{B}(h_i,\epsilon)$时，$\{h_1,\cdots,h_K\}$称为$\Theta$的$\epsilon$-覆盖，即$\forall\theta\in\Theta,\exists i\ s.t.\ d(\theta,h_i)\leq\epsilon$.

定义 3.4 (覆盖数). (Ledoux 等, 2013) 基于定义.3.3，Θ以ϵ为半径的覆盖数可定义为

$$\mathfrak{C}(\epsilon,\Theta,d)=\min\left\{n:\exists\text{大小为}n\text{的}\Theta\text{的}\epsilon-\text{覆盖}\right\}. \tag{3.20}$$

基于上述定义，可得出以下结果。证明见附录 B.5。注意，在以下讨论中，覆盖数是在度量$d_{\infty,\mathcal{S}}$之上定义的。具体而言，给定两个向量值函数 $\boldsymbol{S}=(\boldsymbol{S}^{(1)},\cdots,\boldsymbol{S}^{(N_C)})$、$\tilde{\boldsymbol{s}}=(\tilde{\boldsymbol{s}}^{(1)},\cdots,\tilde{\boldsymbol{s}}^{(N_C)})$，训练数据$\mathcal{S}$，$d_{\infty,\mathcal{S}}(\boldsymbol{S},\tilde{\boldsymbol{s}})$定义为：

$$d_{\infty,\mathcal{S}}(\boldsymbol{S},\tilde{\boldsymbol{s}})=\max_{\boldsymbol{x}_i\in\mathcal{S},j\in[N_C]}|\boldsymbol{S}^{(j)}(\boldsymbol{z})-\tilde{\boldsymbol{s}}^{(j)}(\boldsymbol{z})|. \tag{3.21}$$

定理 3.5 ($\mathrm{MAUC}^{\downarrow}$ Rademacher复杂度的chaining界). 假设得分函数$s^{(i)}$将$\mathcal{X}$ 映射至有界区间$[-R_s,R_s]$，$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})$具有以下性质：

(a) 对于任意单减序列$\{\epsilon_k\}_{k=1}^{\infty}$，若$\lim_{k\to\infty}\epsilon_k=0$且$\epsilon_0\geq R_s$，则：

$$\begin{aligned}\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})&\leq N_C\cdot(N_C-1)\cdot\phi_\ell\cdot\epsilon_K\\&+6\cdot\sum_{k=1}^{K}\epsilon_k\phi_\ell(N_C-1)\cdot\xi(\boldsymbol{Y})\sqrt{\frac{\log(\mathfrak{C}(\epsilon_k,\mathcal{F},d_{\infty,\mathcal{S}}))}{N}}\end{aligned} \tag{3.22}$$

(b) 存在一个常数C，满足：

$$\begin{aligned}\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})&\leq C\phi_\ell\inf_{R_s\geq\alpha\geq0}\Bigg(N_C(N_C-1)\alpha\\&+(N_C-1)\cdot\xi(\boldsymbol{Y})\cdot\int_{\alpha}^{R_s}\sqrt{\frac{\log(\mathfrak{C}(\epsilon,\mathcal{F},d_{\infty,\mathcal{S}}))}{N}}d\epsilon\Bigg)\end{aligned} \tag{3.23}$$

根据定理B.5.4，可进一步通过极小化技术（minorization technique）(Reeve 等, 2020)，构造Rademacher复杂度的简单上界。

定理 3.6 (上界的变形). 给定假设类

$$\mathsf{soft} \circ \mathcal{F} = \left\{ \boldsymbol{g}(\boldsymbol{x}) = \mathsf{soft}(\boldsymbol{s}(\boldsymbol{x})) : \ \boldsymbol{S} \in \mathcal{F} \right\},$$

soft$(\cdot)$是softmax函数。假设$\boldsymbol{S}(x) \in [-R_s, R_s]^{N_C}$和$\ell$是$\phi_\ell$-Lipschitz连续函数，以下不等式成立：

$$\frac{\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathsf{soft} \circ \mathcal{F})}{N_C(N_C-1)} \leq \phi_\ell \left(2^9 \cdot \frac{1}{N_C} \cdot \sqrt{N_C} \cdot \xi(\boldsymbol{Y}) \cdot \log^{3/2}(e \cdot R_s \cdot N \cdot N_C) \cdot \hat{\mathfrak{R}}_{N \cdot N_C}(\Pi \circ \mathcal{F}) + \sqrt{\frac{1}{N}} \right), \tag{3.24}$$

$\hat{\mathfrak{R}}_{N \cdot N_C}(\Pi \circ \mathcal{F})$定义为：

$$\hat{\mathfrak{R}}_{N \cdot N_C}(\Pi \circ \mathcal{F}) = \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f=(f^{(1)},\cdots,f^{(N_C)}) \in \mathcal{F}} \frac{1}{N \cdot N_C} \sum_{j=1}^{N_C} \sum_{i=1}^{N} \sigma_j^{(i)} \cdot f^{(j)}(\boldsymbol{x}_i) \right] \tag{3.25}$$

$\{\sigma_j^{(i)}\}_{(i,j)}$为独立Rademacher随机变量组成的序列。

证明. 基于定理.B.5.4-a)中的chaining界，其证明过程与(Reeve 等, 2020, Prop.1)类似，其中参数为$\lambda = 1,\ \theta = 0,\ q = N_C,\ n = N$.

定理3.6中的结果将成对Rademacher复杂度$\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathsf{soft} \circ \mathcal{F})$转化为样本层次Rademacher复杂度$\hat{\mathfrak{R}}_{N \cdot N_C}$的函数。注意到$\hat{\mathfrak{R}}_{N \cdot N_C}$已不再涉及逐对求和计算，可借鉴标准的Rademacher复杂度上界(Mohri 等, 2018)构造形式方式获得其上界。

3.4.3　深度模型的泛化界

本节进一步基于Rademacher复杂度及其性质，导出三个假设类的泛化界：（1）深层全连接网络，以及（2）深层卷积神经网络。

3.4.3.1　深度全连接网络的泛化界

首先对深层全连接网络进行形式化定义，将具有L个全连接层，N_C个输出单元的深度神经网络表示为：

$$f(\boldsymbol{x}) = \boldsymbol{W} f_{\omega,L}(\boldsymbol{x}) = \boldsymbol{W} s(\boldsymbol{\omega}_{L-2} \cdots s(\boldsymbol{\omega}_1 \boldsymbol{x})),$$

其中：$s(\cdot)$为激活函数；n_{h_j}为第j层的单元数；$\boldsymbol{\omega}_j \in \mathbb{R}^{n_{h_{j+1}} \times n_{h_j}}$，$j = 1, 2, \cdots, L-2$为前$L-1$层权重; $\boldsymbol{W} \in \mathbb{R}^{n_{h_{L-1}} \times N_C}$ 为输出层权重；第i层的输出定义为:

$$f^{(i)}(\boldsymbol{x}) = \boldsymbol{W}^{(i)^\top} f_{\omega,L}(\boldsymbol{x}),$$

$\boldsymbol{W}^{(i)^\top}$为$\boldsymbol{W}$的第$i$行；此外，假定

$$\Pi_{\boldsymbol{W},\omega} = ||\boldsymbol{W}||_F \cdot \prod_{j=1}^{L-2} ||\boldsymbol{\omega}_j||_F \leq \gamma$$

将具有上述特点的深度全连接网络抽象为如下假设类：

$$\begin{aligned}\mathcal{H}^{DNN}_{\gamma,R_s,n_h} = \Big\{ & f : f^{(i)}(\boldsymbol{x}) = \boldsymbol{W}^{(i)^\top} f_{\omega,L}(\boldsymbol{x}), ||f^{(i)}||_\infty \leq R_s, \\ & i = 1, \cdots, N_C,\ \Pi_{\boldsymbol{W},\omega} \leq \gamma \Big\}.\end{aligned} \tag{3.26}$$

为了保证各类别输出具有概率含义，将$f(\boldsymbol{x})$与softmax函数进行复合。最终有效模型可抽象为以下假设类:

$$\mathsf{soft} \circ \mathcal{H}^{DNN}_{\gamma,R_s,n_h} = \left\{ g : g^{(i)} = \frac{\exp(f^{(i)}(\boldsymbol{x}))}{\sum_{j=1}^{N_C} \exp(f^{(j)}(\boldsymbol{x}))}, f \in \mathcal{H}^{DNN}_{\gamma,R_s,n_h} \right\}. \tag{3.27}$$

此类模型有以下泛化界，该结论为附录B.5和引理B.9 中理论结果的自然推演，证明过程基于Talagrand压缩性质，以及定理B.5.4和定理3.6所展现的chaining技术。

定理 3.7 (**深度全连接网络的泛化界**). 基于定理B.5.2中的假设，若：

- $s(\cdot)$为1-Lipshiptz且为正齐次激活函数
- ℓ为ϕ_ℓ-Lipschitz连续函数；
- 输入特征采样自$\mathcal{X} \subset \mathbb{R}^d$，对于所有$\boldsymbol{x} \in \mathcal{X}$，$||\boldsymbol{x}||_2^2 \leq R_{\mathcal{X}}$，

则对于所有$f \in \mathsf{soft} \circ \mathcal{H}^{DNN}_{\gamma,R_s,n_h}$，以下不等式以不低于$1-\delta$的概率成立：

$$R_\ell(f) \leq \hat{R}_S(f) + \min\left(\mathcal{I}_{DNN,1},\ \mathcal{I}_{DNN,2}\right) \cdot \sqrt{\frac{1}{N}} \tag{3.28}$$

$$\chi(\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{N_C} \sum_{j \neq i} \frac{1}{\rho_i \rho_j}},\ \ \xi(\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{N_C} \frac{1}{\rho_i}},\ \ \rho_i = \frac{n_i}{N},$$

$$\mathcal{I}_{DNN,1} = C_1 \frac{\sqrt{2}}{2} \phi_\ell \cdot \frac{\chi(\boldsymbol{Y})}{N_C - 1} + \left(\frac{\sqrt{2} C_1 R_{\mathcal{X}} \phi_\ell \gamma}{2} \cdot C_3 + \frac{C_2 B}{N_C} \cdot \sqrt{2 \log\left(\frac{2}{\delta}\right)} \right) \cdot \xi(\boldsymbol{Y}), \tag{3.29}$$

$$\begin{aligned}\mathcal{I}_{DNN,2} = & C_1 \phi_\ell \left(\frac{2^9}{N_C} \cdot \xi(\boldsymbol{Y}) \cdot \log^{3/2}(K \cdot N \cdot N_C)\, \gamma \cdot R_{\mathcal{X}} \cdot (\sqrt{2\log(2)L} + 1) + 1 \right) \\ & + C_2 \frac{B \cdot \sqrt{\log(\frac{2}{\delta})} \cdot \xi(\boldsymbol{Y})}{N_C},\end{aligned} \tag{3.30}$$

其中 C_1、C_2为常数，$K = e \cdot R_s$，$C_3 = \frac{\sqrt{L \log 2} + \sqrt{N_C}}{\sqrt{N_C - 1}}$。

证明. 由(Golowich 等, 2018)定理.1，易知：$\hat{\mathfrak{R}}_{N\cdot N_C}(\Pi\circ\mathcal{F}) \leq \frac{R_{\mathcal{X}}\gamma(\sqrt{2\log(2)L}+1)}{N\cdot N_C}$ 进一步叠加综合定理.B.5.2，定理.3.6，及附录B.5.5中的引理.B.9，即可完成证明。

其中$\chi(\boldsymbol{Y})$反映了样本中不同类对(i,j)的分布情况，若不同类对分布越为平衡则该变量数值越小。因此$\chi(\boldsymbol{Y})$以二阶的形式反映了数据集的不平衡程度。注意到该结论到泛化误差对网络结构复杂度的依赖仅体现于$\sqrt{L}$即根号下全连接层数。而增加网络复杂度可在对$\sqrt{L}$改变较小的情况下使训练误差$\hat{R}_S(f)$显著降低，由此证明采用深度学习模型的可行性。

3.4.3.2　深度卷积网络的泛化界

进一步在(Long 等, 2020)基础上给出深度卷积网络的MAUC$^{\downarrow}$泛化界。

采用(Long 等, 2020)中的基本设定，考虑如下深度卷积网络假设类。主要研究具有N_{conn}个全连接层和N_{conv}个卷积层的深层神经网络。第i个卷积层的卷积核为线性算子 $\boldsymbol{K}^{(i)} \in \mathbb{R}^{k_i\times k_i\times c_{i-1}\times c_i}$。对于给定的卷积核算子$\boldsymbol{K}$，将卷积操作表示为$op(\boldsymbol{K})$，$\boldsymbol{K}(\boldsymbol{x}) = op(\boldsymbol{K})\boldsymbol{x}$。此外，假设每个卷积层之后都连接非线性激活函数以及一个可选的池化操作，并假设激活函数和池化操作都都满足1-Lipschitz连续性质。对于第i个全连接层，将其权重表示为 $\boldsymbol{V}^{(i)}$。综合卷积层和全连接层，给定的深度神经网络的所有参数集合可表示为 $\boldsymbol{P} = \{\boldsymbol{K}^{(1)},\cdots,\boldsymbol{K}^{(N_{conv})},\boldsymbol{V}^{(1)},\cdots,\boldsymbol{V}^{(N_{conn})}\}$。同样地，假设损失函数满足：

1. ϕ_ℓ-Lipschitz
2. $Range\{\ell\} \subseteq [0,B]$

在此设定下给出所研究的假设类。首先定义$\mathcal{P}_\nu^{(0)}$为模型初始参数的假设类：

$$\begin{aligned}\mathcal{P}_\nu^{(0)} = \Bigg\{\boldsymbol{P} : &\left(\max_{i\in\{1,\cdots,N_{conv}\}} ||op(\boldsymbol{K}^{(i)})||_2\right) \\ &\leq 1+\nu, \left(\max_{j\in\{1,\cdots,N_{conn}\}} ||\boldsymbol{V}^{(j)}||_2\right) \leq 1+\nu\Bigg\}.\end{aligned} \tag{3.31}$$

进一步地，将所学模型限定于以下$\mathcal{P}_{\beta,\nu}$假设类中。参数与$\mathcal{P}_\nu^{(0)}$中的固定初始值之间距离不大于β：

$$\mathcal{P}_{\beta,\nu} = \left\{\boldsymbol{P} : d_{NN}(\boldsymbol{P},\tilde{\boldsymbol{P}}_0) \leq \beta,\ \tilde{\boldsymbol{P}}_0 \in \mathcal{P}_\nu^{(0)}\right\}. \tag{3.32}$$

其中，$d_{NN}(\cdot,\cdot)$给出了两组参数的距离：

$$d_{NN}(\boldsymbol{P},\tilde{\boldsymbol{P}}) = \sum_{i=1}^{N_{conv}} ||op(\boldsymbol{K}^{(i)}) - op(\tilde{\boldsymbol{K}}^{(i)})||_2 + \sum_{i=1}^{N_{conn}} ||\boldsymbol{V}^{(i)} - \tilde{\boldsymbol{V}}^{(i)}||_2. \tag{3.33}$$

基于上述定义，应用定理B.5.4可得出以下结论，证明过程见引理B.10。

定理 3.8 (**深度卷积网络的泛化界**). 若采用softmax函数作为网络输出层进行处理，给定如下深度卷积网络假设类：

$$\text{soft} \circ \mathcal{F}_{\beta,\nu} = \left\{ \boldsymbol{g}(\boldsymbol{x}) = \text{soft}(\boldsymbol{s}_{\boldsymbol{P}}(\boldsymbol{x})) : \ \boldsymbol{s}_{\boldsymbol{P}} \in \mathcal{F}_{\beta,\nu} \right\},$$

$$\begin{aligned} \mathcal{F}_{\beta,\nu} = \{ s_{\boldsymbol{P}} : \mathbb{R}^{N_{N_L-1}} \to \mathbb{R}^{N_C} \mid \boldsymbol{P} \in \mathcal{P}_{\beta,\nu}, \\ Range(\boldsymbol{s}_P) \subseteq [-R_s, R_s]^{N_C} \}. \end{aligned} \tag{3.34}$$

此外，假设

- $\sup_{x \in \mathcal{X}} ||vec(\boldsymbol{x})|| \leq R_{\mathcal{X}}$；
- $R_s > 1/\min \left\{ \sqrt{N}, \frac{\xi(\boldsymbol{Y})}{N_C} \cdot \sqrt{N_{par} \left(\nu N_L + \beta + \log(3R_{\mathcal{X}} \cdot \beta \cdot N) \right)} \right\}$；
- 给定数据集$\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$，各样本由独立采样得到。
- 替代损失函数ℓ的值域包含于$[0, B]$,

则对于所有得分函数 $f \in \text{soft} \circ \mathcal{F}_{\beta,\nu}$， $\forall \delta \in (0,1)$时，以下不等式以不低于$1-\delta$的概率成立:

$$R_\ell(f) \leq \hat{R}_{\mathcal{S}}(f) + C_1 \alpha_1 \alpha_2 + C_2 \cdot \frac{B\xi(\boldsymbol{Y})}{N_C} \cdot \sqrt{\frac{\log(\frac{2}{\delta})}{N}}, \tag{3.35}$$

$$\begin{aligned} \alpha_1 &= \tilde{C} \cdot \frac{\phi_\ell \cdot R_s \cdot \xi(\boldsymbol{Y})}{N_C}, \\ \alpha_2 &= \sqrt{\frac{N_{par} \left(\nu N_L + \beta + \log\left(3\beta R_{\mathcal{X}} N\right) \right)}{N}}, \end{aligned}$$

$\tilde{C}, C_1, C_2$为常数， $N_L = N_{conv} + N_{conn}$， N_{par}是神经网络的参数总量。

证明. 由定理.B.5.2及附录B.5.5中的引理.B.11即可得证。

3.4.3.3 小结

注意，以上泛化上界中具有两个不平衡感知因子。第一个因子$\xi(\boldsymbol{Y})$可捕获$\mathcal{S}$中的类标签分布不平衡性。另一个因子$\chi(\boldsymbol{Y})$可捕获$\mathcal{S}$中类别对分布的不平衡性。因此，为提升泛化能力，训练集$\mathcal{S}$必须同时具有较大的样本数N和较小的不平衡性。其中不平衡性可由$(\xi(\boldsymbol{Y}), \chi(\boldsymbol{Y}))$感知。换言之，该泛化界表明：盲目

增加训练数据集的样本量并不能提高泛化能力。仅当增加少数类的样本点，才可从根源上解决泛化瓶颈。因此，与常见的$O(\sqrt{1/N})$结果相比，本章的泛化界能更好地刻画训练数据中的不平衡问题。

3.5　优化加速

在本节中，重点讨论优化过程中存在的实际问题。如之前几节所示，MAUC$^{\downarrow}$替代损失的复杂形式为下游优化算法造成较大的计算负担。具体地，给定单个样本上计算损失的时间复杂度T_ℓ以及梯度计算的时间复杂度T_{grad}，不难看出，即使一次全批量或小批量损失和梯度计算，也需要 $O(\sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j T_\ell)$和$O(\sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j T_{grad})$的复杂度，其量级接近平方规模，难以获得较好的拓展性。然而，本章发现对于一些主流的替代损失，成对计算过程可大幅简化。具体地，提出指数损失、平方损失和铰链损失函数的损失及梯度的加速计算方法，复杂度可大致由平方规模下降至线性规模。

首先考察目标函数及其梯度计算的一般过程。具体而言，经验替代风险函数$\hat{R}_\ell$具有以下一般形式：

$$\hat{R}_\ell = \sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_i n_j}\cdot \ell^{i,j}, \tag{3.36}$$

其中，

$$\ell^{i,j} = \ell\left(f^{(i)}(\boldsymbol{x}_m) - f^{(i)}(\boldsymbol{x}_n)\right),$$

$$f^{(i)}(\boldsymbol{x}) = g_i(\boldsymbol{W}h_{\boldsymbol{\theta}}(\boldsymbol{x})),$$

$$\boldsymbol{W} = [\boldsymbol{w}^{(1)},\cdots,\boldsymbol{w}^{(N_C)}]^\top.$$

如，当将 $g_i(\cdot)$定义为神经网络最后一层（如softmax函数）的激活函数时，$\boldsymbol{w}^{(i)}$即可视为输出层的权重，$h_{\boldsymbol{\theta}}(\cdot)$是除去最后输出层外的神经网络， $f^{(i)}(\boldsymbol{x})$则为整体神经网络。注意到，$g_i(\cdot)$和$h_{\boldsymbol{\theta}}(\cdot)$的选择只会影响样本层次的chaining法则，而仅ℓ的计算与平方规模复杂度相关。

首先给定通用计算过程。对于输出层参数，假设$\boldsymbol{w}^{(i)}\in\mathbb{R}^{n_h\times 1}$，将$\boldsymbol{w}^{(i)}$逐列进行拼接，得到矩阵$\boldsymbol{W}$，即，

$$\boldsymbol{W} = [\boldsymbol{w}^{(1)},\cdots,\boldsymbol{w}^{(N_C)}]^\top,\ \ \boldsymbol{W}\in\mathbb{R}^{N_C\times n_h}. \tag{3.37}$$

注意到向量化操作并不影响求导过程，为方便起见，将模型除$\boldsymbol{W}$外的参数进行拼接并进行向量化，得到参数向量$\boldsymbol{\theta} \in \mathbb{R}^{d_\theta \times 1}$。例如，在深度学习网络中，$\boldsymbol{\theta}$是除输出层参数外所有参数的拼接。另外，$h_{\boldsymbol{\theta}}(\boldsymbol{x})$ 为一个向量，满足$h_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^{n_h \times 1}$，$\boldsymbol{H}$ 是每个实例对应的$h_{\boldsymbol{\theta}}(\boldsymbol{x})$ 的拼接，即：

$$\boldsymbol{H} = [h_{\boldsymbol{\theta}}(\boldsymbol{x}_1), \cdots, h_{\boldsymbol{\theta}}(\boldsymbol{x}_N)]^\top, \ \boldsymbol{H} \in \mathbb{R}^{n_h \times N}. \tag{3.38}$$

推导加速算法过程中需要使用中间导数的计算，相关变量定义如下：

$$\partial_i f^{(j)}(\boldsymbol{x}) = \frac{\partial (f^{(j)}(\boldsymbol{x}))}{\partial \boldsymbol{w}^{(i)\top} h_{\boldsymbol{\theta}}(\boldsymbol{x})} \tag{3.39}$$

$$\begin{aligned} &\boldsymbol{\partial}_{i,j} = [\partial_i f^{(j)}(\boldsymbol{x}_1), \cdots, \partial_i f^{(j)}(\boldsymbol{x}_N)]^\top, \ \boldsymbol{\partial}_{i,j} \in \mathbb{R}^{N \times 1}, \\ &\boldsymbol{\partial}_j(\boldsymbol{x}_k) = [\partial_1 [f^{(j)}(\boldsymbol{x}_k)], \cdots, \partial_{N_C}(f^{(j)}(\boldsymbol{x}_k))]^\top, \ \boldsymbol{\partial}_j(\boldsymbol{x}_k) \in \mathbb{R}^{N_C \times 1}. \end{aligned} \tag{3.40}$$

应用链式法则时，需要如下变量：

$$\boldsymbol{U}^{(j)} = [\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(\boldsymbol{x}_1) \boldsymbol{W}^\top \boldsymbol{\partial}_j(\boldsymbol{x}_1), \cdots, \nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(\boldsymbol{x}_N) \boldsymbol{W}^\top \boldsymbol{\partial}_j(\boldsymbol{x}_N)], \ \boldsymbol{U}^{(j)} \in \mathbb{R}^{d_\theta \times N}. \tag{3.41}$$

最后，记权重$\frac{1}{n_i n_j}$的向量形式为：

$$\boldsymbol{D}^{(i)} = [\frac{1}{n_i n_{y_1}}, \cdots, \frac{1}{n_i n_{y_N}}]^\top, \ \boldsymbol{D}^{(i)} \in \mathbb{R}^{N \times 1}. \tag{3.42}$$

3.5.1 指数损失

3.5.1.1 损失计算

对于指数损失，可以通过因式分解进行简化计算：

$$\begin{aligned} \hat{R}_{exp} &= \sum_{i=1}^{N_C} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \cdot \exp\left(\alpha \cdot \left(f^{(i)}(\boldsymbol{x}_m) - f^{(i)}(\boldsymbol{x}_n)\right)\right), \\ &= \sum_{i=1}^{N_C} \underbrace{\left(\sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \exp(\alpha \cdot f^{(i)}(\boldsymbol{x}_m))\right)}_{(a_i)} \cdot \underbrace{\left(\sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \cdot \exp(-\alpha \cdot f^{(i)}(\boldsymbol{x}_n))\right)}_{(b_i)} \end{aligned} \tag{3.43}$$

基于上述变形，可先分别计算(a_i)、(b_i)，然后将其相乘，从而进行损失计算。相比较原始的$O(\sum_{i=1}^{N_C} \sum_{j \neq i} n_i n_j N_C T_\ell)$复杂度，上述变形只需$O(N N_C T_\ell)$复杂度，大大提高了计算效率。

3.5.1.2　梯度计算

对梯度计算的加速可由类似方式得出。由公式(3.37)-公式(3.41)及链式法则，有：

$$
\begin{aligned}
\nabla_{\boldsymbol{w}^{(i)}}(a_j) &= \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \alpha \exp(\alpha \cdot f^{(j)}(\boldsymbol{x}_m)) \cdot \partial_i f^{(j)}(\boldsymbol{x}_m) \cdot h_{\boldsymbol{\theta}}(\boldsymbol{x}_m) \\
&= \alpha \boldsymbol{H}(\mathcal{I}_i \odot \hat{\boldsymbol{y}}_{exp}^{(i),+} \odot \boldsymbol{\partial}_{i,j}). \\
\nabla_{\boldsymbol{w}^{(i)}}(b_j) &= -\sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \alpha \cdot \exp(-\alpha \cdot f^{(j)}(\boldsymbol{x}_n)) \cdot \partial_i f^{(j)}(\boldsymbol{x}_n) \cdot h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) \\
&= -\alpha \boldsymbol{H}(\mathcal{I}_{\setminus i} \odot \boldsymbol{D}^{(i)} \odot \hat{\boldsymbol{y}}_{exp}^{(i),-} \odot \boldsymbol{\partial}_{i,j}), \\
\nabla_{h_{\boldsymbol{\theta}}(\boldsymbol{x}_k)}(a_j) &= \underbrace{\mathcal{I}_j(\boldsymbol{x}_k) \cdot \alpha \cdot \exp(\alpha \cdot f^{(j)}(\boldsymbol{x}_k))}_{q_j(\boldsymbol{x}_k)} \cdot \sum_{i=1}^{N_C} \partial_i f^{(j)}(\boldsymbol{x}_k) \cdot \boldsymbol{w}^{(i)}, \\
&= q_j(\boldsymbol{x}_k) \cdot \boldsymbol{W}^\top \boldsymbol{\partial}_j(\boldsymbol{x}_k) \\
\nabla_{h_{\boldsymbol{\theta}}(\boldsymbol{x}_k)}(b_j) &= -\underbrace{\mathcal{I}_j^c(\boldsymbol{x}_k) \cdot \boldsymbol{D}_k^{(i)} \cdot \alpha \cdot \exp(-\alpha \cdot f^{(j)}(\boldsymbol{x}_k))}_{\tilde{q}_j(\boldsymbol{x}_k)} \cdot \sum_{i=1}^{N_C} \partial_i f^{(j)}(\boldsymbol{x}_k) \cdot \boldsymbol{w}^{(i)}, \\
&= -\tilde{q}_j(\boldsymbol{x}_k) \cdot \boldsymbol{W}^\top \boldsymbol{\partial}_j(\boldsymbol{x}_k)
\end{aligned}
\tag{3.44}
$$

其中，

$$
\begin{aligned}
&\mathcal{I}_i = I\left[\boldsymbol{x}_i \in \mathcal{N}_i\right],\ \mathcal{I}_{\setminus i} = I\left[\boldsymbol{x}_i \notin \mathcal{N}_i\right],\ \mathcal{I}_j(\boldsymbol{x}) = \boldsymbol{I}\left[x \in \mathcal{N}_j\right],\ \mathcal{I}_j^c(\boldsymbol{x}) = 1 - \mathcal{I}_j(\boldsymbol{x}) \\
&\hat{\boldsymbol{y}}_{exp}^{(i),+} = [\exp(\alpha \cdot f^{(i)}(\boldsymbol{x}_1)), \cdots, \exp(\alpha \cdot f^{(i)}(\boldsymbol{x}_N))]^\top, \\
&\hat{\boldsymbol{y}}_{exp}^{(i),-} = [\exp(-\alpha \cdot f^{(i)}(\boldsymbol{x}_1)), \cdots, \exp(-\alpha \cdot f^{(i)}(\boldsymbol{x}_N))]^\top.
\end{aligned}
\tag{3.45}
$$

此外，定义$\boldsymbol{q}_j = [q_j(\boldsymbol{x}_1), \cdots, q_j(\boldsymbol{x}_N)]^\top$，$\tilde{\boldsymbol{q}}_j = [\tilde{q}_j(\boldsymbol{x}_1), \cdots, \tilde{q}_j(\boldsymbol{x}_N)]^\top$。根据公式(3.41)，有：

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}}(a_j) &= \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_k)) \nabla_{h_{\boldsymbol{\theta}}(\boldsymbol{x}_k)} a_j = \boldsymbol{U}^{(j)} \boldsymbol{q}_j \\
\nabla_{\boldsymbol{\theta}}(b_j) &= \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_k)) \nabla_{h_{\boldsymbol{\theta}}(\boldsymbol{x}_k)} b_j = -\boldsymbol{U}^{(j)} \tilde{\boldsymbol{q}}_j
\end{aligned}
\tag{3.46}
$$

算法 3 CalIndex

输入：$\boldsymbol{x}_0^{\downarrow}, \cdots, \boldsymbol{x}_{n_i-1}^{\downarrow}, \tilde{\boldsymbol{x}}_0^{\downarrow}, \cdots, \tilde{\boldsymbol{x}}_{N-n_i-1}^{\downarrow}, \alpha$.
$p \leftarrow 0$ ▷ 当前正实例的指针
$q \leftarrow 0$ ▷ 当前负实例的指针
$\mathcal{I}_{0:(n_i-1)} = 0$. ▷ 满足$\mathcal{A}_j^{(i)} = \{\tilde{\boldsymbol{x}}_k^{(i)\downarrow}\}_{k=1}^{\mathcal{I}_k}$的指示集
while $p < n_i, q < N - n_i$ **do**
  if $f^{(i)}(\boldsymbol{x}_p^{(i)\downarrow}) - f^{(i)}(\tilde{\boldsymbol{x}}_q^{(i)\downarrow}) < \alpha$ **then**
    $q+ = 1$ ▷ 若当前实例对非零，转到下一个负实例
  else
    $\mathcal{I}_p = \max(q-1, 0)$ ▷ $\mathcal{A}_p^{(i)}$ 为$\tilde{\boldsymbol{x}}_{q-1}^{(i)\downarrow}$的最后一个元素
    $p+ = 1$
  end if
end while
if $q = N - n_i$ **then**
  $\mathcal{I}_p = \max(q-1, 0)$ ▷ $\mathcal{A}_p^{(i)}$ 为$\tilde{\boldsymbol{x}}_{N-n_i-1}^{(i)\downarrow}$的最后一个元素
end if
if $p < n_i$ **then** ▷ 若$\mathcal{A}_p^{(i)}$ 包括所有负实例，则对于所有$\mathcal{A}_{p'}^{(i)} = \mathcal{A}_p^{(i)}$，$p' > p$
  $\mathcal{I}_{(p+1):(n_i-1)} = \mathcal{I}_p$
end if

从而，可通过下式进行梯度计算：

$$\begin{aligned}
\nabla_{\boldsymbol{w}^{(i)}} \hat{R}_{surr} &= \sum_{j=1}^{N_C} \left(\nabla_{\boldsymbol{w}^{(i)}}(a_j)\right) \cdot (b_j) + (a_j) \cdot \left(\nabla_{\boldsymbol{w}^{(i)}}(b_j)\right) \\
&= \alpha \boldsymbol{H} \left(\sum_{j=1}^{N_C} b_j \mathcal{I}_i \odot \hat{\boldsymbol{y}}_{exp}^{(i),+} \odot \boldsymbol{\partial}_{i,j} - a_j \mathcal{I}_{\backslash i} \odot \boldsymbol{D}^{(i)} \odot \hat{\boldsymbol{y}}_{exp}^{(i),-} \odot \boldsymbol{\partial}_{i,j} \right), \\
\nabla_{\boldsymbol{\theta}} \hat{R}_{surr} &= \sum_{j=1}^{N_C} \left(\nabla_{\boldsymbol{\theta}}(a_j)\right) \cdot (b_j) + (a_j) \cdot \left(\nabla_{\boldsymbol{\theta}}(b_j)\right) \\
&= \sum_{j=1}^{N_C} \left[\boldsymbol{U}^{(j)} (b_j \boldsymbol{q}_j - a_j \tilde{\boldsymbol{q}}_j) \right].
\end{aligned} \tag{3.47}$$

由上述紧凑形式可见，相比较原始的$O(N_C \sum_{i=1}^{N_C} \sum_{j \neq i} n_i n_j T_{grad})$复杂度，加速算法仅需$O(N N_C T_{grad})$ 复杂度。

3.5.2 铰链损失

3.5.2.1 损失计算

首先给出铰链替代风险的形式：

$$\hat{R}_{hinge} = \sum_{i=1}^{N_C} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \cdot \left(\alpha - \left(f^{(i)}(\boldsymbol{x}_m) - f^{(i)}(\boldsymbol{x}_n) \right) \right)_+ . \tag{3.48}$$

注意到，求和内部的项仅当$\left(f^{(i)}(\boldsymbol{x}_m) - f^{(i)}(\boldsymbol{x}_n)\right) \leq \alpha$时非零。此外，对于非零

项，$\max(x,0)=x$。一旦获得铰链损失的所有非零项，损失计算则退化为自反函数，后续计算可显著加速。因此，加速算法的关键在于高效地定位非零项。

固定一个类别i和一个实例$\boldsymbol{x}_m \in \mathcal{N}_i$，记$\mathcal{A}^{(i)}(\boldsymbol{x}_m)$为与$\boldsymbol{x}_m$关联的所有非零项集合：

$$\mathcal{A}^{(i)}(\boldsymbol{x}_m)=\{\boldsymbol{x}_n \notin \mathcal{N}_i, \alpha > f^{(i)}(\boldsymbol{x}_m)-f^{(i)}(\boldsymbol{x}_n)\}.$$

根据$\mathcal{A}^{(i)}(\boldsymbol{x}_m)$，可将$\hat{R}_{hinge}$重形式化为：

$$\begin{aligned}\hat{R}_{hinge} &= \sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\left[\left(\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j}\right)\cdot\left(\alpha-f^{(i)}(\boldsymbol{x}_m)\right)\right.\\ &\quad\left.+\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j}\cdot f^{(i)}(\boldsymbol{x}_n)\right],\\ &= \sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\left[\delta^{(i)}(\boldsymbol{x}_m)\cdot(\alpha-f^{(i)}(\boldsymbol{x}_m))+\Delta^{(i)}(\boldsymbol{x}_m)\right],\end{aligned} \tag{3.49}$$

其中，

$$\delta^{(i)}(\boldsymbol{x}_m)=\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j},\ \ \Delta^{(i)}(\boldsymbol{x}_m)=\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j}f^{(i)}(\boldsymbol{x}_n).$$

基于该形式，一旦$\delta^{(i)}(\boldsymbol{x})$ 和$\Delta^{(i)}(\boldsymbol{x})$被固定，可在$O(N)$的复杂度内完成对损失函数的计算。故计算瓶颈主要来自于对$\delta^{(i)}(\boldsymbol{x})$和$\Delta^{(i)}(\boldsymbol{x})$的计算。

以下介绍高效计算$\delta^{(i)}(\boldsymbol{x}_m),\Delta^{(i)}(\boldsymbol{x}_m),\mathcal{A}^{(i)}(\boldsymbol{x}_m)$的方法。为此，给定一个具体的类别$i$，首先根据实例得分$f^{(i)}(\boldsymbol{x})$对其正负例样本分别进行排序：

$$\begin{aligned}&f^{(i)}(\boldsymbol{x}_0^{\downarrow})\geq f^{(i)}(\boldsymbol{x}_1^{\downarrow})\cdots,\geq f^{(i)}(\boldsymbol{x}_{n_i-1}^{\downarrow}),\quad \boldsymbol{x}_i^{\downarrow}\in\mathcal{N}_i,\\ &f^{(i)}(\tilde{\boldsymbol{x}}_0^{\downarrow})\geq f^{(i)}(\tilde{\boldsymbol{x}}_1^{\downarrow})\cdots,\geq f^{(i)}(\tilde{\boldsymbol{x}}_{N-n_i-1}^{\downarrow}),\quad \tilde{\boldsymbol{x}}_i^{\downarrow}\notin\mathcal{N}_i.\end{aligned} \tag{3.50}$$

显然，易知：

$$\mathcal{A}^{(i)}(\boldsymbol{x}_0^{\downarrow})\subseteq\mathcal{A}^{(i)}(\boldsymbol{x}_1^{\downarrow})\subseteq\cdots\subseteq\mathcal{A}^{(i)}(\boldsymbol{x}_{n_i-1}^{\downarrow}). \tag{3.51}$$

从而，利用动态规划算法，仅需对全体样本进行一次扫描，即可找出所有$\mathcal{A}^{(i)}(\boldsymbol{x}_k^{\downarrow}),k=0,1,\cdots,n_i-1$，具体方法如算法3。基于构建好的$\mathcal{A}^{(i)}(\boldsymbol{x}_k^{\downarrow})$（为简化表示，将其记作$\mathcal{A}_k^{(i)}$），给出如下递归计算$\delta^{(i)}(\boldsymbol{x})$和$\Delta^{(i)}(\boldsymbol{x})$的方式：

$$\begin{aligned}\delta^{(i)}(\boldsymbol{x}_{k+1}^{\downarrow})&=\delta^{(i)}(\boldsymbol{x}_k^{\downarrow})+\sum_{j\neq i}\sum_{\mathcal{N}_j\cap\mathcal{A}_{k+1}^{(i)}\setminus\mathcal{A}_k^{(i)}}\frac{1}{n_i n_j},\\ \Delta^{(i)}(\boldsymbol{x}_{k+1}^{\downarrow})&=\Delta^{(i)}(\boldsymbol{x}_k^{\downarrow})+\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}_{k+1}^{(i)}\setminus\mathcal{A}_k^{(i)}}\frac{1}{n_i n_j}f^{(i)}(\boldsymbol{x}_n).\end{aligned} \tag{3.52}$$

算法 4 铰链损失加速算法

输入：$\boldsymbol{X}$、$\boldsymbol{Y}$、α。
输出：loss。
loss $\leftarrow 0$
for $i = 1 : N_C$ **do**
  基于$f^{(i)}(\boldsymbol{x})$计算$\boldsymbol{x}_0^{(i)\downarrow}, \cdots, \boldsymbol{x}_{n_i-1}^{(i)\downarrow}$
  基于$f^{(i)}(\boldsymbol{x})$计算$\boldsymbol{x}_0^{(i)\downarrow}, \cdots, \boldsymbol{x}_{N-n_i-1}^{(i)\downarrow}$
  $f^{(i)}(\boldsymbol{x})_+ = [f^{(i)}(\boldsymbol{x}_0^{(i)\downarrow}), \cdots, f^{(i)}(\boldsymbol{x}_{n_i-1}^{(i)\downarrow})]$
  $\boldsymbol{D}^{(i)} = \left[\frac{1}{n_i n_{y(\tilde{\boldsymbol{x}}_0^{(i)\downarrow})}}, \cdots, \frac{1}{n_i n_{y(\tilde{\boldsymbol{x}}_{N-1}^{(i)\downarrow})}}\right]$
  $t_0 \leftarrow 0,\ \ \delta_{\text{offset}}^{(i)} \leftarrow 0,\ \Delta_{\text{offset}}^{(i)} \leftarrow 0,\ \ loc = 0$
  $\boldsymbol{\Delta}_{0:(n_i-1)}^{(i)} \leftarrow 0,\ \ \boldsymbol{\delta}_{0:(n_i-1)}^{(i)} \leftarrow 0$
  $\mathcal{I} \leftarrow$ CalIndex$(\boldsymbol{x}_0^{(i)\downarrow}, \cdots, \boldsymbol{x}_{n_i}^{(i)\downarrow}, \tilde{\boldsymbol{x}}_0^{(i)\downarrow}, \cdots, \tilde{\boldsymbol{x}}_{N-n_i-1}^{(i)\downarrow}, \alpha)$

  while $k < n_i$ **do**
    if $\mathcal{I}_k \neq t_0 - 1$ **then** ▷ 当$\mathcal{A}^{(k)} \neq \mathcal{A}^{(k-1)}$时，继续扫描
      $t_1 \leftarrow \mathcal{I}_k + 1$ ▷ 确保$\mathcal{A}_{k+1}^{(i)} \backslash \mathcal{A}_k^{(i)} = \{\tilde{\boldsymbol{x}}_u^{(i)\downarrow}\}_{u=t_0}^{t_1-1}$
      $\delta_k^{(i)} \leftarrow \delta_{\text{offset}}^{(i)} + \sum_{u=t_0}^{t_1-1} \boldsymbol{D}_u^{(i)}$ ▷ 递归计算$\delta^{(i)}(\boldsymbol{x}_k^{(i)\downarrow})$
      $\Delta_k^{(i)} \leftarrow \Delta_{\text{offset}}^{(i)} + \sum_{u=t_0}^{t_1-1} \boldsymbol{D}_u^{(i)} \cdot f^{(i)}(\tilde{\boldsymbol{x}}_k^{(i)\downarrow})$ ▷ 递归计算$\Delta^{(i)}(\boldsymbol{x}_k^{(i)\downarrow})$
      $t_0 \leftarrow t_1$ ▷ 更新最初的元素
      $\delta_{\text{offset}}^{(i)} \leftarrow \delta_k^{(i)},\ \ \Delta_{\text{offset}}^{(i)} \leftarrow \Delta_k^{(i)}$ ▷ 更新偏置
    else if $I_k \neq N - n_i - 1$ **then** ▷ 检查是否满足$\mathcal{A}^{(k)} = \mathcal{A}^{(k-1)}$且$\mathcal{A}^{(k)} \neq \mathcal{N}_{\backslash i}$
      $\delta_k^{(i)} = \delta_{\text{offset}}^{(i)},\ \ \Delta_k^{(i)} = \Delta_{\text{offset}}^{(i)}$ ▷ 复制最后一次更新
    else
      break ▷ 当$\mathcal{A}^{(k)} = \mathcal{N}_{\backslash i}$且$\mathcal{A}^{(k)} = \mathcal{A}^{(k-1)}$时停止迭代
    end if
    k+ = 1
  end while
  if $k < n_i$ **then**
    $\boldsymbol{\delta}_{k:(n_i-1)}^{(i)} = \delta_{\text{offset}}^{(i)},\ \ \boldsymbol{\Delta}_{k:(n_i-1)}^{(i)} = \Delta_{\text{offset}}^{(i)}$ ▷ 令$\mathcal{A}_x^{(i)} = \mathcal{A}_{k-1}^{(i)},\ \forall x \geq k$
  end if
  loss+ = $(\alpha - f^{(i)}(\boldsymbol{x})_+)^\top \boldsymbol{\delta}^{(i)} + \boldsymbol{\Delta}^{(i)}\mathbf{1}$
end for

上述递归规则可同样由线性时间实现。综上所述，铰链损失加速算法仅需$O(N_C\bar{N}T_l)$的时间复杂度。算法4给出所述加速算法的具体实现步骤。

3.5.2.2 梯度计算

与损失计算类似，给出梯度计算的加速算法。由类似分析可知：

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}}\hat{R}_{hinge} &= \sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\left[-\left(\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j}\right)\cdot\nabla_{\boldsymbol{\Theta}}f^{(i)}(\boldsymbol{x}_m)\right.\\ &\quad\left.+\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j}\nabla_{\boldsymbol{\Theta}}f^{(i)}(\boldsymbol{x}_n)\right] \\ &= \sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\left(-\delta^{(i)}(\boldsymbol{x}_m)\cdot\nabla_{\boldsymbol{\Theta}}f^{(i)}(\boldsymbol{x}_m)+\Gamma^{(i)}(\boldsymbol{x}_m)\right)\end{aligned} \tag{3.53}$$

其中，

$$\Gamma^{(i)}(\boldsymbol{x}_m)=\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}^{(i)}(\boldsymbol{x}_m)}\frac{1}{n_i n_j}\nabla_{\boldsymbol{\Theta}}f^{(i)}(\boldsymbol{x}_n).$$

对于任意$\boldsymbol{x}\in\mathcal{S}$，任意类别$j$，基于公式(3.37)-公式(3.41)，可通过下式计算$\nabla_{\Theta}f^{(j)}(\boldsymbol{x})$：

$$\begin{aligned}&\nabla_{\boldsymbol{w}^{(i)}}f^{(j)}(\boldsymbol{x})=\partial_i f^{(j)}(\boldsymbol{x})\cdot h_{\boldsymbol{\theta}}(\boldsymbol{x}),\\ &\nabla_{h_{\boldsymbol{\theta}}(\boldsymbol{x})}f^{(i)}(\boldsymbol{x})=\boldsymbol{W}^{\top}\boldsymbol{\partial}_j(\boldsymbol{x}),\\ &\nabla_{\boldsymbol{\theta}}f^{(i)}(\boldsymbol{x})=\nabla_{\boldsymbol{\theta}}h_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{W}^{\top}\boldsymbol{\partial}_j(\boldsymbol{x}).\end{aligned} \tag{3.54}$$

同损失计算类似，采用动态规划算法进行加速。第一步得到$\boldsymbol{x}_0^{\downarrow},\cdots,\boldsymbol{x}_{n_i-1}^{\downarrow}$和$\tilde{\boldsymbol{x}}_0^{\downarrow},\cdots,\tilde{\boldsymbol{x}}_{N-n_i-1}^{\downarrow}$，并计算 $\mathcal{A}_k^{(i)}$（过程与损失计算相同）。随后，采用如下递归方式计算新的变量$\Gamma^{(i)}(\boldsymbol{x}_k^{\downarrow})$：

$$\Gamma^{(i)}(\boldsymbol{x}_{k+1}^{\downarrow})=\Gamma^{(i)}(\boldsymbol{x}_k^{\downarrow})+\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j\cap\mathcal{A}_{k+1}^{(i)}\setminus\mathcal{A}_k^{(i)}}\frac{1}{n_i n_j}\nabla_{\boldsymbol{\Theta}}f^{(i)}(\boldsymbol{x}_n). \tag{3.55}$$

基于上式，可采用类似算法进行梯度计算。

3.5.3　平方损失

3.5.3.1　损失计算

对于每个固定的类别i，构造一个相似度矩阵$\boldsymbol{Aff}^{(i)}$，满足：

$$\boldsymbol{Aff}_{m,n}^{(i)}=\begin{cases}\dfrac{1}{n_i n_{y_m}}, & n\in\mathcal{N}_i, m\notin\mathcal{N}_i,\\ \dfrac{1}{n_i n_{y_n}}, & m\in\mathcal{N}_i, n\notin\mathcal{N}_i,\\ 0, & \text{otherwise.}\end{cases} \tag{3.56}$$

从而，有矩阵形式：

$$\boldsymbol{Aff}^{(i)} = \boldsymbol{D}^{(i)}(\mathbf{1} - \mathbf{Y}^{(i)})\mathbf{Y}^{(i)\top} + \mathbf{Y}^{(i)}(\mathbf{1} - \mathbf{Y}^{(i)})^{\top}\boldsymbol{D}^{(i)}, \tag{3.57}$$

随后，平方替代损失下的经验替代风险函数可重形式化为：

$$\hat{R}_{sq} = \sum_{i=1}^{N_C} \boldsymbol{\Delta}_{sq}^{(i)\top} \mathcal{L}^{(i)} \boldsymbol{\Delta}_{sq}^{(i)}, \tag{3.58}$$

其中，

$$\boldsymbol{\Delta}_{sq}^{(i)} = \boldsymbol{Y}^{(i)} - f^{(i)}(\boldsymbol{X}), \;\; f^{(i)}(\boldsymbol{X}) = [f^{(i)}(\boldsymbol{x}_1), \cdots, f^{(i)}(\boldsymbol{x}_N)]^{\top}, \tag{3.59}$$

且$\mathcal{L}^{(i)} = diag(\boldsymbol{Aff}^{(i)}\mathbf{1}) - \boldsymbol{Aff}^{(i)}$ 是$\boldsymbol{Aff}^{(i)}$对应的拉普拉斯矩阵。观察$\mathcal{L}^{(i)}$的结构，可推出一个损失加速计算的高效分解方案。具体而言，方案如下：

$$\hat{R}_{sq} = \sum_{i=1}^{N_C} \frac{1}{2} \boldsymbol{\Delta}_{sq}^{(i)\top} (\boldsymbol{\kappa}^{(i)} \odot \boldsymbol{\Delta}_{sq}^{(i)}) - \Delta_1^{(i)} \cdot \Delta_2^{(i)}, \tag{3.60}$$

其中，

$$\begin{aligned} \Delta_2^{(i)} &= \boldsymbol{Y}^{(i)\top} \boldsymbol{\Delta}_{sq}^{(i)} \\ \Delta_1^{(i)} &= \boldsymbol{\Delta}_{sq}^{(i)\top} \boldsymbol{D}^{(i)} (\mathbf{1} - \boldsymbol{Y}^{(i)}) \\ \boldsymbol{\kappa}^{(i)} &= n_i \boldsymbol{D}^{(i)} (\mathbf{1} - \boldsymbol{Y}^{(i)}) + \frac{N_C - 1}{n_i} \boldsymbol{Y}^{(i)}. \end{aligned} \tag{3.61}$$

显然，上式包含的两项均只需$O(N_C N T_l)$的复杂度。具体分解方案见算法5。

3.5.3.2 梯度计算

根据算法5, 公式(3.37)-公式(3.41)和链式法则，有：

$$\nabla_{\boldsymbol{w}^{(i)}} \hat{R}_{sq} = \sum_{i=1}^{N_C} \frac{1}{2} \cdot \nabla_{\boldsymbol{w}^{(i)}} \boldsymbol{\Delta}_{sq}^{(i)\top} (\boldsymbol{\kappa}^{(i)} \odot \boldsymbol{\Delta}_{sq}^{(i)}) - \nabla_{\boldsymbol{w}^{(i)}} \Delta_1^{(i)} \cdot \Delta_2^{(i)} \tag{3.62}$$

其中，

$$\begin{aligned} \frac{1}{2} \nabla_{\boldsymbol{w}^{(i)}} \boldsymbol{\Delta}_{sq}^{(j)\top} (\boldsymbol{\kappa}^{(j)} \odot \boldsymbol{\Delta}_{sq}^{(j)}) &= \boldsymbol{H}(\boldsymbol{\Delta}_{sq}^{(j)} \odot \boldsymbol{\kappa}^{(j)} \odot \boldsymbol{\partial}_{i,j}), \\ \nabla_{\boldsymbol{w}^{(i)}} \Delta_1^{(i)} \Delta_2^{(i)} &= \boldsymbol{H} \left[\left(\Delta_2^{(i)} \cdot \boldsymbol{D}^{(i)} (\mathbf{1} - \boldsymbol{Y}^{(i)}) + \Delta_1^{(i)} \boldsymbol{Y}^{(i)} \right) \odot \boldsymbol{\partial}_{i,j} \right], \\ \frac{1}{2} \nabla_{\boldsymbol{\theta}} \boldsymbol{\Delta}_{sq}^{(j)\top} (\boldsymbol{\kappa}^{(j)} \odot \boldsymbol{\Delta}_{sq}^{(j)}) &= \boldsymbol{U}^{(j)} (\kappa^{(j)} \odot \boldsymbol{\Delta}^{(j)}), \\ \nabla_{\boldsymbol{\theta}} \Delta_1^{(i)} \Delta_2^{(i)} &= \boldsymbol{U}^{(j)} \left[\Delta_2^{(i)} \cdot \boldsymbol{D}^{(i)} (\mathbf{1} - \boldsymbol{Y}^{(i)}) + \Delta_1^{(i)} \boldsymbol{Y}^{(i)} \right], \end{aligned} \tag{3.63}$$

该结论再次说明梯度计算仅需$O(N_C N T_{grad})$ 时间复杂度。

算法 5 平方损失加速算法

输入: $\boldsymbol{X}$、$\boldsymbol{Y}$、α、$\varphi(\cdot)$.

输出: loss。

$\texttt{loss} \leftarrow 0$

for $i = 1 : N_C$ **do**

　　计算$\boldsymbol{D}^{(i)}$。

　　$\boldsymbol{\Delta}^{(i)} \leftarrow (\boldsymbol{Y}^{(i)} - f^{(i)}(\boldsymbol{X}))$

　　$\boldsymbol{\kappa}^{(i)} \leftarrow n_i \boldsymbol{D}^{(i)}(\mathbf{1} - \boldsymbol{Y}^{(i)}) + \dfrac{N_C - 1}{n_i}\boldsymbol{Y}^{(i)}$

　　$\Delta_1^{(i)} \leftarrow \left(\boldsymbol{\Delta}_{sq}^{(i)\top}(\boldsymbol{D}^{(i)}\mathbf{1} - \boldsymbol{Y}^{(i)})\right)$

　　$\Delta_2^{(i)} \leftarrow \boldsymbol{Y}^{(i)\top}\boldsymbol{\Delta}_{sq}^{(i)}$

　　$\texttt{loss+} = \frac{1}{2}\boldsymbol{\Delta}_{sq}^{(i)\top}(\boldsymbol{\kappa}^{(i)} \odot \boldsymbol{\Delta}_{sq}^{(i)}) - \Delta_1^{(i)} \cdot \Delta_2^{(i)}$

end for

表 3.1 三个替代损失函数的加速

Table 3.1 Acceleration for three losses[2]

算法	损失函数	梯度	要求
exp + acceleration	$O(N_C \cdot N \cdot T_\ell)$	$O(N_C \cdot N \cdot T_{grad})$	$\min_i n_i \gg 2$
squared + acceleration	$O(N_C \cdot N \cdot T_\ell)$	$O(N_C \cdot N \cdot T_{grad})$	$e^{\frac{1}{2}(N-n_i)} \gg n_i \gg N - e^{\frac{1}{2}n_i}$
hinge + acceleration	$O(N_C \cdot \bar{N} \cdot T_\ell)$	$O(N_C \cdot \bar{N} \cdot T_{grad})$	$\min_i n_i \gg 2$
w/o acceleration	$O(\sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j \cdot T_\ell)$	$O(\sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j \cdot T_{grad})$	\

3.5.4　总结

表3.1比较了不同算法的复杂度。为保证足够的效率提升，数据集大小必须满足$N_C \cdot N < \sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j$且$N_C \cdot \bar{N} < \sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j$。首先，注意到：

$$\sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j - N_C \cdot N = \sum_{i=1}^{N_C}\left[n_i(N-n_i) - (n_i + (N-n_i))\right]. \tag{3.64}$$

为保证$N_C \cdot N < \sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j$，需令

$$n_i(N-n_i) > (n_i + (N-n_i)), \quad \forall i.$$

[2] $\bar{N} = \sum_{i=1}^{N_C} n_i \log n_i + (N-n_i)\log(N-n_i)$

显然，当$\min_i n_i \gg 2$时满足要求。类似地，有：

$$\sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j - N_C \cdot \bar{N} = \sum_{i=1}^{N_C} [n_i(N-n_i) - n_i \log(n_i) - (N-n_i)\log(N-n_i)] \,. \quad (3.65)$$

从而，当

$$\frac{1}{2}(N-n_i) > \log(n_i), \quad \frac{1}{2}(n_i) > \log(N-n_i)$$

时，满足$N_C \cdot \bar{N} < \sum_{i=1}^{N_C}\sum_{j\neq i} n_i n_j, \ \forall i$。换言之，应保证满足

$$\exp(\frac{1}{2}(N-n_i)) \gg n_i \gg N - \exp(\frac{1}{2}n_i), \ \forall i$$

即可得到明显的效率提升。且上述要求除极端不平衡的数据外，大部分真实数据集均可满足。

每类样本数量

1 ~ 10	100	50	360	20	20	300	420	20	100	200
11 ~ 20	100	100	600	50	100	50	50	100	100	100
21 ~ 30	50	100	600	100	50	600	100	20	200	20
31 ~ 40	600	100	100	300	50	100	20	600	100	600
41 ~ 50	20	600	100	300	20	300	420	600	480	100
51 ~ 60	20	20	50	100	20	420	600	50	20	100
61 ~ 70	50	100	200	360	100	100	100	50	100	50
71 ~ 80	50	150	600	50	300	20	150	20	20	100
81 ~ 90	20	480	50	50	100	20	50	20	300	50
91 ~ 100	50	360	100	150	20	100	100	20	100	50

图 3.1 CIFAR-100-Imb数据集标签分布，表格中的每个单元代表一个特定的n_i。左侧的行标题显示行中id的范围。类id的编号方式与原始数据集相同。

Figure 3.1 The label distribution for CIFAR-100-Imb Dataset, where each cell in the table presents a specific n_i. The row title in the left shows the range of id in the row. The class ids are numbered in the same way as the original dataset.

3.6 实验

3.6.1 数据集

本节主要对数据集进行实证分析。首先介绍所使用数据集，总体而言，数据集有三种来源：（a）LIBSVM网站，（b）KEEL网站，及（c）其他网站。注意，

对于所有以Imb为前缀的数据集，本章使用从原始数据集中采样的不平衡子集进行实验。数据集的基本信息摘要如表.3.3所示。 r_χ定义为 $r_\chi = \frac{\max_i n_i}{\min_i n_i}$ (Lemaître等, 2017)。

表 3.2 User-Imb数据集的n_i

Table 3.2 n_i for User-Imb

class	F22-	F24-26	F27-28	F29-32	F33-42	F43+
sample	800	400	400	800	1,600	400
class	M22-	M23-26	M27-28	M29-31	M32-38	M39+
sample	1,600	8,000	800	1,600	8,000	8,000

表 3.3 数据集基本信息

Table 3.3 Basic Information of the Datasets

Dataset	$\#samples$	$\#classes$	$\#features$	r_χ
Balance	625	3	4	5.88
Dermatology	357	6	34	5.55
Ecoli	336	8	7	71.50
New-thyroid	215	3	5	5.00
Pageblocks	548	5	10	164.00
SegmentImb	749	7	18	20.31
Shuttle	2,175	5	9	853.00
Svmguide2	391	3	20	4.17
Yeast	1,484	10	8	92.60
CIFAR-100-Imb	23,350	100	2,048	50.00
User-Imb	24,400	12	21,527	20.00

(a) **LIBSVM数据集 [3]:** 包括**Shuttle**, **Svmguide-2**, **SegmentImb**。

(b) **KEEL数据集 [4]:** 包括**Balance**, **Dermatology**, **Ecoli**, **New Thyroid**, **Page Blocks**, **Yeast**。

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[4] http://www.keel.es/

(c) **其他数据集：**：

(1) **CIFAR-100-Imb：** CIFAR-100数据集[5]包含来自100类的600个样本。本章通过采样形成CIFAR-100数据集的不平衡版本。100个类编码为$1, \cdots 100$。对于每个类i，随机抽取如图3.1中所示的n_i实例，形成数据集的不平衡版本。

(2) **User-Imb：** 原始数据集来自第三方移动数据平台TalkingData，该平台根据移动用户的应用程序使用记录预测移动用户的人口统计特征。该数据集是为Kaggle比赛而收集，名为Talking Data Mobile User Demographics。[6] 原始特征包括记录的事件、应用程序属性和设备信息等。有12个目标类'F23-'、'F24-26'、'F27-28'、'F29-32'、 'F33-42'、'F43+'、'M22-'、'M23-26'、'M27-28'、'M29-31'、'M32-38'、 'M39+'。这些类别描述用户的人口统计信息（性别和年龄）。在本章的实验中，对原始数据集的不平衡子集进行采样。采样后的各类别样本数如表.3.2所示。

3.6.2 对比方法

对比方法的选择： 对比方法包括以下几类：第一类为**基线模型**不针对不平衡问题进行任何处理；第二类为基于采样的方法，通过**过采样和欠采样**处理不平衡问题；第三类为**不平衡损失函数**：通过特定损失函数处理不平衡问题。此外，通过平方损失、指数损失和铰链损失以及第3.5节中提出的加速方法实现本章的框架。其中采样操作通过python库 `Imbalanced learn`(Lemaître 等, 2017) [7] 进行实现。具体情况如下：

1. **基线模型：Logistics 回归 (LR)**. 对于传统的数据集，将**LR**损失应用于线性模型。对于深度学习数据集，将交叉熵损失应用于公共的网络骨架。
2. **过采样方法：** 首先增加少数类样本的比例，生成更平衡的数据集，然后使用**LR**在新数据集上训练模型。使用的采样策略如下：

    - **BM：** (BorderlineSMOTE) (Han 等, 2005): 该方法为SMOTE方法的一个变种，通过对位于决策边界的少数难样本进行重复采样，从而增加难样本的比例。

    - **MM：** (MWMOTE) (Barua 等, 2012): 该方法为SMOTE方法的另一个变

[5] https://www.cs.toronto.edu/~kriz/cifar.html

[6] https://www.kaggle.com/c/talkingdata-mobile-user-demographics/data

[7] https://imbalanced-learn.readthedocs.io/en/stable/#

种。该方法首先识别出较难的少数类样本，并根据这些样本与最近的多数类样本之间的欧氏距离确定权重。

3. **欠采样方法**. 首先通过采样减少多数类样本数，生成一个更均衡的数据集，然后使用**LR**在新的数据集上训练模型。采用以下三种对比方法：

   • **IHT**： (InstanceHardnessThreshold) (Smith 等, 2014). 该方法利用(Smith 等, 2014)中提出的实例难度度量从多数类中过滤难样本和噪声样本。

   • **NM**： (NearMiss)(Mani 等, 2003): 该方法采用欠采样的思想，使多数类样本包围少数类样本。

   • **TL**： (TomekLinks) (Tomek, 1976) 该方法采用Tomek-Links方法去除多数类中对决策边界贡献不大的冗余样本。

4. **不平衡损失函数** (应用于深度学习模型)：对于CIFAR-100-Imb和 User-Imb数据集，将本章所提方法与不平衡损失函数进行对比。

   • **Focal Loss**(Lin 等, 2017a): 该类方法通过在交叉熵损失中加入一个调节因子来突出训练过程中的难样本和少数样本，从而解决不平衡问题。

   • **CB-CE**: 该方法是指在交叉熵损失上应用(Cui 等, 2019)中提出的样本加权方案解决不平衡问题。

   • **CB-Focal**: 该方法是指在Focal损失上应用(Cui 等, 2019)中提出的样本加权方案解决不平衡问题。

   • **LDAM**(Cao 等, 2019): 在各类别最小边际的基础上，该方法提出可感知标签分布的边际损失。

5. 所提出的方法：

   • **Ours1**： 基于平方替代损失$\ell_{sq}(\alpha,t)=(\alpha-t)^2$实现本章所提框架。

   • **Ours2**： 基于指数损失$\ell_{\exp}(\alpha,t)=\exp(-\alpha t)$实现本章所提框架。

   • **Ours3**： 基于铰链损失$\ell_{hinge}(\alpha,t)=(\alpha-t)_+$实现本章所提框架。

3.6.3　实现细节

基本架构： 所有实验均在一台配备Intel(R) Xeon(R) CPU E5-2620 v4 cpu和TITAN RTX GPU的ubuntu16.04.6服务器上进行。实验代码使用`python3.6.7`实现，依赖库包括： `pytorch` (v-1.1.0)、`sklearn` (v-0.21.3)和`numpy` (v-1.16.2)。对于传统的数据集，借助`sklearn`和`numpy`实现所提算法。对于铰链损失，使用Cython来加速动态规划算法。对于深度学习数据集，所提算法使用`pytorch`实

现。

度量指标：给定得分函数$f=(f^{(1)},\cdots,f^{(N_C)})$，通过MAUC指标进行性能评估，MAUC越大代表性能越好：

$$\begin{aligned}\text{MAUC}^{\uparrow}=&\frac{1}{N_C\cdot(N_C-1)}\cdot\\&\sum_{i=1}^{N_C}\sum_{j\neq i}\frac{\left|\left\{(\boldsymbol{x}_1,\boldsymbol{x}_2):\boldsymbol{x}_1\in\mathcal{N}_i,\ \boldsymbol{x}_2\in\mathcal{N}_j,f^{(i)}(\boldsymbol{x}_1)>f^{(i)}(\boldsymbol{x}_2)\right\}\right|}{n_in_j}\end{aligned}\tag{3.66}$$

优化算法：在优化方法上，采用类nestrov加速法进行非凸优化；采用(Li 等, 2015b; Liu 等, 2019a,b)中的算法训练softmax线性模型；采用adam算法训练深层神经网络。此外，采用SGD算法训练User-Imb数据集的对应模型；使用adam(Zou 等, 2019; Kingma 等, 2015)训练CIFAR-100-Imb数据集的对应模型。

传统数据集：所有实验的超参数均通过训练集和验证集进行调参，通过测试集上的结果对其进行评估。其中，训练集、验证集和测试集分别占原始数据集的80%, 10%, 10%。为保留相同标签分布，根据不同类别进行分层采样生成训练集、验证集和测试集。对每个算法独立进行了15次重复实验。鉴于所有传统数据集均只包括简单特征，采用线性模型加softmax变换作为得分函数。形式化表述为：给定一个实例$\boldsymbol{x}$，有$f(\boldsymbol{x})=\text{softmax}(\boldsymbol{W}\boldsymbol{x})$，其中，$\boldsymbol{W}=[\boldsymbol{\omega}^{(1)},\cdots,\boldsymbol{\omega}^{(N_C)}]\in\mathbb{R}^{d\times N_C}$是线性函数的权重。针对基于采样的对比方法，首先对数据集进行相应采样，然后基于多分类交叉熵损失训练得分函数。对于LR基线模型，直接基于多分类交叉熵损失训练得分函数。对于所提出算法，直接基于$\text{MAUC}^{\downarrow}$替代损失和得分函数f直接进行训练。所提出算法采用的超参数如表3.4所示。

深度学习数据集：所有实验的超参数均通过训练集和验证集进行调参，通过测试集上的结果对其进行评估。其中，训练集、验证集和测试集分别占原始数据集的80%, 10%, 10%。为保留相同标签分布，根据不同类别进行分层采样生成训练集、验证集和测试集。对每个算法独立进行了15次重复实验。对于两个深度学习数据集，所有模型均基于一个深度神经网络。得分函数均形如$f(\boldsymbol{x})=\text{softmax}(fc_2(fc_1(\boldsymbol{x})))$，其中，$fc_1,fc_2$是全连接层。对于CIFAR-100-Imb数据集，将图片输入ResNet-50预训练模型得到的`layer4`经过池化层后的特征作为每个方法的输入特征。对于User-Imb数据集，从头训练模型。针对基于采样的对比方法，首先对数据集进行相应采样，然后基于多分类交叉熵损失同时训练得分函

表 3.4 传统数据集上的最优参数设定(λ, α)

Table 3.4 Best parameters for (λ, α) over traditional datasets.

数据集	Ours1	Ours2	Ours3
balance	(1e-4, 0.9)	(1e-4, 0.9)	(1e-4, 0.7)
dermatology	(1e-4, 0.9)	(1e-4, 0.9)	(1e-4, 0.3)
ecoli	(6e-4, 0.6)	(4e-4, 0.9)	(1e-4, 0.2)
new-thyroid	(1e-4, 0.9)	(1e-4, 0.5)	(1e-4, 0.9)
pageblocks	(6e-4, 0.8)	(2e-4, 0.9)	(6e-4, 0.5)
segmentImb	(2e-4, 0.9)	(4e-4, 0.9)	(1e-4, 0.3)
shuttle	(2e-4, 0.8)	(1e-4, 0.7)	(1e-4, 0.5)
svmguide2	(1e-4, 0.9)	(1e-4, 0.9)	(2e-4, 0.4)
yeast	(6e-4, 0.3)	(9e-3, 0.1)	(1e-4, 0.2)

数和三层神经网络。对于LR基线模型，直接基于多分类交叉熵损失训练得分函数。所提出算法在CIFAR-100-Imb数据集和User-Imb数据集上采用的超参数分别如表3.5和表3.6所示。

表 3.5 CIFAR-100-Imb数据集上的超参数设定

Table 3.5 Hyperparameters for CIFAR-100-Imb

	批大小	学习率	权重衰减	学习率衰减	gamma	学习率衰减间隔（epo.）
Ours1	1000	$1.2E-3$	$5E-6$	0.97	1.00	5.00
Ours2	1000	$1E-3$	$1E-6$	0.99	4.00	5.00
Ours3	1000	$1E-3$	$5E-5$	0.99	4.00	3.00

表 3.6 User-Imb数据集上的超参数设定

Table 3.6 Hyperparameters for User-Imb

	批大小	学习率	权重衰减	学习率衰减	gamma	学习率衰减间隔（epo.）
Ours1	32	0.005	.0001	0.97	0.5	1.00
Ours2	32	0.005	.0001	0.97	2.0	1.00
Ours3	32	0.005	.0001	0.97	2.0	1.00

3.6.4 实验结果

加速算法的数值验证： 首先验证第3.5节所加速算法的有效性。注意，加速算法只涉及损失和梯度的计算，而不依赖于所采用的优化算法。因此，为公平起见，首先独立地测试未经加速算法的运行时间。具体地，随机生成一组数据集 $\boldsymbol{Z}_i = (\boldsymbol{X}_i, \boldsymbol{Y}_i),\ i = 1, 2, \cdots, 5,\ \ \boldsymbol{X}_i \in \mathbb{R}^{N_k \times d}, \boldsymbol{Y}_i \in \mathbb{R}^{N_i \times N_C}$ 是输入特征矩阵和所输出的one-hot标签矩阵。具体地，令$N_i = \{32, 64, 128, 256, 512, 1024\},\ d = 100,\ N_C = 5$。 $\boldsymbol{X}_i$由$[0, 1]$范围内的均匀分布生成。记$\rho_i = \frac{n_i}{N}$，随机生成$\boldsymbol{Y}$使得满足下述条件：$\rho_1 = 0.2, \rho_2 = 0.1, \rho_3 = 0.2, \rho_4 = 0.4, \rho_5 = 0.1$。此外，使用softmax修正$f(\cdot)$的输出，即$f^{(i)}(\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}^{(i)}\boldsymbol{x})$。权重矩阵$\boldsymbol{W} = [\boldsymbol{W}^{(1)}, \cdots, \boldsymbol{W}^{(5)}]$的生成方式与 $\boldsymbol{X}$相同。基于此，分别在不同的 $\boldsymbol{Z}_i$上测试一般计算方法和加速计算方法的运行时间。 30次重复实验的加速比均值如图3.2所示。其中加速比定义

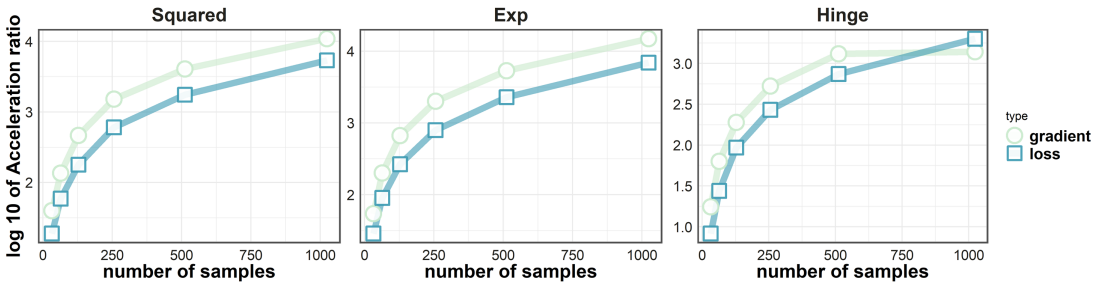


图 3.2 加速比 vs. 样本规模

Figure 3.2 Acceleration Ratio vs. Sample Scale

为：

$$\text{加速比} = \frac{\text{一般方法运行时间}}{\text{加速方法运行时间}}. \tag{3.67}$$

由图可知：（a）所提加速算法在所有测试数据集上均取得显著加速效果。当样本数量小于32时，加速比在10 ~ 100范围内，随着样本数量增大，结果逐渐提升，最终当样本数达到1024时，平方损失和指数损失的加速达到10000倍。该结果表明所提算法可同时适用于小批量和全批量的优化算法。（b）加速比曲线大致呈现$O(N)$趋势，该结果与表.3.1中的复杂度分析一致。具体地，对于平方损失和指数损失：

$$\begin{aligned} \text{加速比} &\approx \frac{\sum_{i=1}^{N_C} \sum_{j \neq i} n_i n_j}{N_C \cdot N} \\ &= \frac{\sum_{i=1}^{N_C} \sum_{j \neq i} \rho_i \cdot \rho_j}{N_C} \cdot N \end{aligned} \tag{3.68}$$

注意，实验设置ρ_i和N_C为常数，因此加速比为$O(N)$。铰链损失虽引入了额外因子$\log n_i$和$\log(N-n_i)$，但由于测试样本数适中，这些因子的值仍保持在常数级别。因此，铰链损失的加速度比呈线性趋势。此外，当样本数超过500时，计算铰链损失梯度的加速比增长趋缓。该现象的原因在于铰链损失的部分计算通过Cython实现。对于大型数据集，Cython实现对损失的作用比对梯度的作用更明显。具体而言，梯度计算以矩阵运算为主，即使不使用Cython也可利用numpy实现加速，而损失计算（主要包含初等运算）中使用Cython则加速效果较为明显。

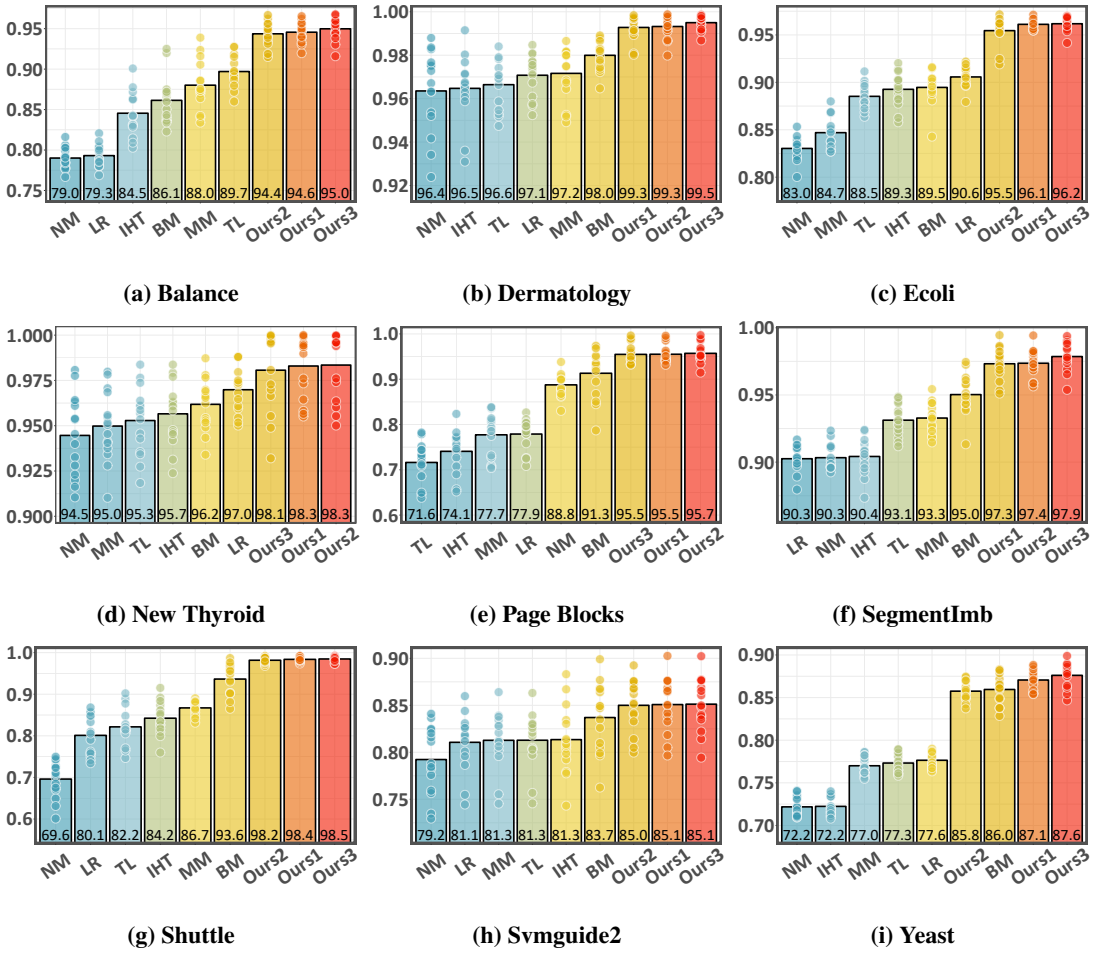


图 3.3 粗粒度性能对比

Figure 3.3 Coarse-grained Performance Comparison

传统数据集：9个传统数据集上15次重复实验的平均性能如图.3.3所示，其中散点为15次不同数据划分对应的观测值，条形图的高代表15次实验的平均性能。由图可知：1）在所有的数据集上，所提加速算法的始终显著优于其他对比方法。具体地，Ours1、Ours2和Ours3中最高性能与其他方法相比在*Balance*, *Der-*

matology, *Ecoli*, *New Thyroid*, *Page Blocks*, *SegmentImb*, *Shuttle*, *Svmguide2*和*Yeast*数据集上，分别提升5.3, 1.5, 5.6, 1.3, 4.4, 2.9, 4.9, 1.4, 1.8。事实证明，所提方法在多数情况下都有显著提升。 2）由观察可知，部分采样方法的性能没有超过LR算法。原因可能在于采样方法无法直接优化AUC。 3）由于不平衡数据集上的性能瓶颈来自其少数类，本章进一步研究少数类别对与性能的关系。具体地，将频率最低5个类别对的性能可视化，如图.3.4。注意，本章只给出大于0.7的结果，以便更清楚地展示最优算法之间的差异。结果表明，所提算法在少数类别对上提升更为显著，在 Balance, Ecoli, Page Blocks, SegmentImb, Shuttle,和 Yeast数据集的提升尤为明显。

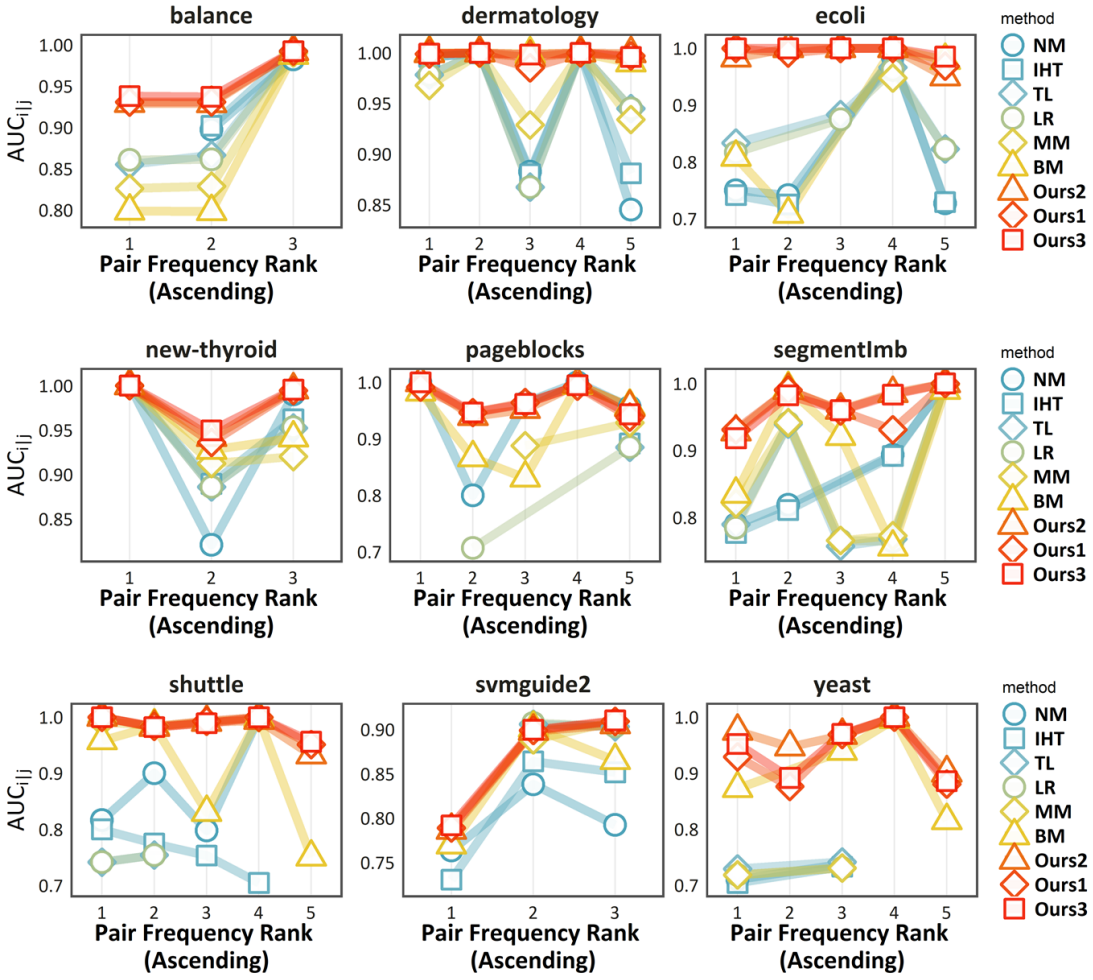


图 **3.4** 稀有类别对的细粒度对比（传统数据集）

Figure 3.4 Fine-grained Comparison Over the Minority Class Pairs (Traditional Datasets)

深度学习数据集：CIFAR-100-Imb和User-Imb的性能对比如表.3.7。观察可知，当使用平方损失（Ours1）或指数损失（Ours2）时，所提算法在所有的数

表 3.7 基于深度学习MAUC↑性能对比

Table 3.7 Performance comparison based on MAUC↑ with Deep Learning

	method	CIFAR-100-Imb	User-Imb
Baseline	CE	59.80	55.26
Sampling based methods	BM	59.07	58.95
	MM	58.52	55.35
	IHT	60.56	55.87
	NM	60.24	54.52
	TL	60.28	52.99
Imbalanced loss	FOCAL	60.03	55.34
	CBCE	60.04	58.91
	CBFOCAL	60.42	58.75
	LDAM	60.26	56.47
Ours	Ours1	61.90	60.64
	Ours2	**62.08**	**60.94**
	Ours3	59.64	59.52

据集上均优于其他方法。此外，5个少数类别对的性能对比如表.3.8所示。对于CIFAR-100-Imb数据集，除第4个类别对外，Ours1, Ours2和Ours3取得的最佳性能在所有类别对上均显著优于其他对比方法。

此外，即使铰链损失（Ours3）未能带来总体MAUC↑性能的提升，但该方法仍在前四个少数类别对上取得了更优的性能，即，缓解了不平衡问题。对于User-Imb，所提方法性能提升较不明显。可能原因在于：（1）User-Imb的不平衡程度相对较低；（2）该数据集中，即使少数类样本量也高达400个。上述两点缓解了数据集的不平衡问题，从而使得各个方法间的性能差异不显著。

[8]在两个深度学习数据集的少数类别对上提供更细粒度的比较结果。1st, 2nd, 3rd, 4th, 5th分别为$p_i p_j$值最小的类别对(i, j)

表 3.8 少数类别对的细粒度对比 (深度学习模型)

Table 3.8 Fine-grained Comparison Over the Minority Class Pairs (Deep Learning Datasets)[8]

type	method	CIFAR-100-Imb					User-Imb				
	bottom pairs	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
Baseline	CE	48.82	53.28	50.74	58.44	56.17	49.43	49.51	53.72	58.06	49.73
Sampling Based methods	BM	50.46	50.65	51.97	54.44	56.79	54.64	53.04	56.36	58.44	58.50
	MM	46.45	38.70	39.07	52.61	44.71	52.77	49.68	54.82	60.42	48.39
	IHT	46.18	51.60	53.53	53.67	58.63	56.49	54.77	58.02	61.60	52.00
	NM	49.01	55.40	54.35	**60.78**	61.75	49.38	51.21	53.99	56.05	52.31
	TL	48.75	52.88	50.43	52.11	55.83	47.55	47.96	50.11	52.35	50.92
Imbalanced loss	FOCAL	51.25	52.58	48.89	58.17	60.38	49.83	49.88	54.16	58.86	49.77
	CBCE	48.42	53.33	52.53	55.06	58.67	**61.48**	**59.05**	61.60	64.42	**58.82**
	CBFOCAL	49.74	51.18	49.45	58.22	58.29	61.45	58.98	61.39	64.09	58.75
	LDAM	48.68	50.45	52.57	57.17	59.96	50.46	52.18	55.20	60.64	49.33
Ours	Ours1	51.18	57.48	59.39	60.33	62.04	58.74	58.20	62.41	64.49	57.58
	Ours2	52.57	54.68	57.31	59.33	**65.38**	59.02	58.54	**62.90**	**65.26**	57.55
	Ours3	**54.34**	**60.70**	**61.18**	58.17	65.13	59.52	56.59	60.09	63.56	56.37

3.7 小结

本章主要探索如何将AUC诱导的机器学习方法应用于较为复杂的多类问题。具体地，本章提出一个新的框架，该框架通过优化M度量来学习得分函数。在优化过程中，使用0-1损失的可微替代损失函数作为优化目标。通过一致性分析，给出M度量下达到fisher一致的充分条件。此外，本章提出一个经验替代风险最小化框架，在保证泛化上界的前提下最小化$\mathrm{MAUC}^{\downarrow}$。在实际应用中，本章还为该框架的三种实现构造加速计算方法。最后，在11个数据集上的实验证明所提方法的有效性。

第4章　基于层级化分解的多任务AUC优化方法及应用

4.1　引言

本节首先对视觉属性及个性化属性学习进行简要介绍，随后介绍本文提出的多任务AUC优化方法。

视觉属性（Visual attributes）是指描述诸如纹理、颜色、情绪等视觉特征的语义线索。典型案例包括鞋子是否舒适、是否为高跟鞋，人物表情是微笑或是哭泣。过去十年中，视觉属性学习已逐渐成为大量应用构建的重要基石(Song 等, 2014; Su 等, 2017; Wang 等, 2017; Yang 等, 2018)。

现有视觉属性学习方法主要基于聚合自少量标注者的全局标签(Farhadi 等, 2009; Sadovnik 等, 2013; Luo 等, 2018)。最近，以Amazon Mechanical Turk为代表的在线众包平台逐渐兴起，使得从大量标注者中获取属性标注成为可能(Kovashka 等, 2015)，并为视觉属性学习提供了进一步发展空间。传统视觉属性学习方法默认用户标注具有共识性，其基本假设为用户决策仅会随机地轻微偏离共识。而粒度理解中，不同标注者可能对属性含义的理解可能不同，例如“开放的”和“时尚的”。该现象表明不可简单地通过随机噪声建模个人决策和共识之间的差距。在更极端情况下，不同用户的标注结果甚至互相矛盾。因此，当个性化标注可知时，有必要同时学习用户共识和个性化观点带来的影响，从而对个性化标注进行预测。针对该问题，本文主要考虑一下两大要点：

首先，介于全体共识和个体决策之间的群体效应（group effect）对于理解特定于用户的属性标注结果存在重要作用。如前人所述(Kovashka 等, 2015)，人们通常根据文化背景和理解词汇语义的方式形成各种“思想流派”。尽管个性化会导致不同流派或群体之间观点迥异，但同一群体内很有可能得到相似的决策。此外，由于群体之间往往存在显著差异，因而不同群体可能喜欢不同的视觉线索。换言之，应当赋予每个群体不同的特征子集，而视觉特征和用户均应同时划分至不同群体以保证性能。鉴于无法事先获得用户特征分组，因此有必要通过模型参数的结构隐式地实现该约束。

其次，与基于共识的属性预测不同，个性化属性学习中偏好学习比标签预测更重要，正如推荐或图片检索。此时，若将属性词汇视为关键词或标签，

应保证标记为正的实例排名高于标记为负的实例。众所周知，ROC曲线下面积（AUC）符合上述要求(Yang 等, 2017)，因此是该任务更合适的优化目标。

本章旨在基于上述两个问题学习个性化属性偏好。具体而言，将每位用户的属性偏好学习均视为一项任务，并在此基础上针对所解决的问题提出多任务模型。本章主要贡献包括：a）在多任务模型中，提出任务参数的层级化分解，即主流用户共识、用户群体聚类和个性化三层要素。b）采用近端梯度下降法求解模型参数，为群体要素的近端算子推导出闭式解，并进一步设计一种基于AUC的加速计算方法。c）对方法收敛性和泛化能力进行系统的理论分析，同时应用于一个仿真数据集和两个真实属性标注数据上。理论和实验结果均显示所提出方法的优越性。

4.2 方法形式化

本节提出了一种基于AUC的多任务模型。以下首先介绍本章使用的符号，以及问题设定和多级参数分解。之后，系统地阐述所提出模型的两个模块：基于AUC的损失和计算方法和正则化算子。

4.2.1 符号

$\langle\cdot,\cdot\rangle$表示矩阵或向量内积。矩阵$\boldsymbol{A}$的奇异值表示为 $\sigma_1(\boldsymbol{A}),\cdots,\sigma_m(\boldsymbol{A})$，且有$\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots, \geq \sigma_m(\boldsymbol{A}) \geq 0$。$\boldsymbol{I}$为单位矩阵。$I[\mathcal{A}]$用以指示集合$\mathcal{A}$，$\mathbf{1}$表示全一向量或矩阵。$\mathcal{U}(a,b)$和$\mathcal{N}(\mu,\sigma^2)$分别代表均匀分布和正态分布。$\otimes$为笛卡尔积。

4.2.2 问题设定

给定某一属性[1]，假设存在U名参与标注图片的用户，本章假设第i名用户标注了 n_i幅图片，其中包括$n_{+,i}$个正标签和$n_{-,i}$个负标签。用户i的正、负样本集合可表示为$\mathcal{S}_{+,i} = \{k \mid y_k^{(i)} = 1\}$和$\mathcal{S}_{-,i} = \{k \mid y_k^{(i)} = -1\}$，而训练集则可表示为$\mathcal{S} = \left\{(\boldsymbol{X}^{(1)},\boldsymbol{y}^{(1)}),\cdots,(\boldsymbol{X}^{(U)},\boldsymbol{y}^{(U)})\right\}$。对于训练集$\mathcal{S}$，$\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_i \times d}$表示第$i$名用户所标注图片的输入特征。$\boldsymbol{X}^{(i)}$的每一行均表示由对应图片提取的特征，$\boldsymbol{y}^{(i)} \in \{-1,1\}^{n_i}$为对应的标签向量。如果$y_k^{(i)} = 1$，则表示用户认为第$k$幅图片具有对应的属性，反之则有$y_k^{(i)} = -1$。

[1]本章所述模型分别学习不同属性，因此以下讨论聚焦于单个特定属性。

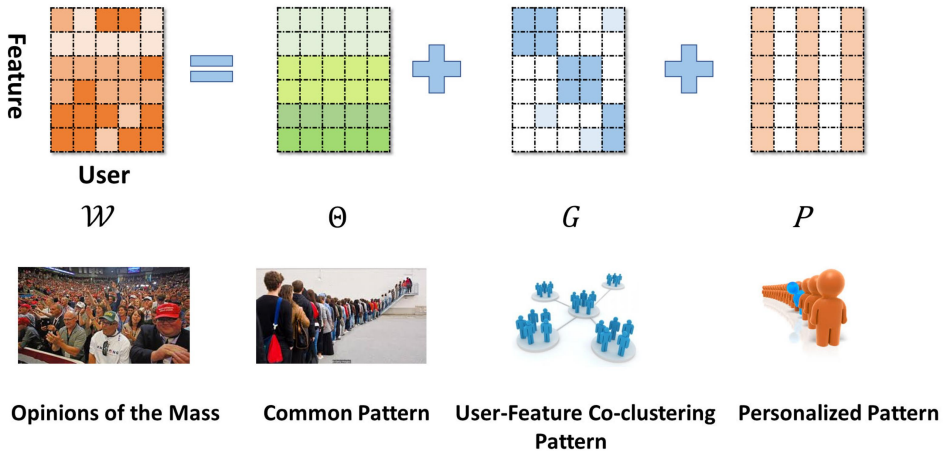


图 4.1 模型参数层级化分解的示意图

Figure 4.1 An illustration of the Multi-level Decomposition of the model parameters

依据多任务学习范式，本章将每位用户的属性偏好学习视为一项任务，并旨在学习所有任务对应的模型$\boldsymbol{f}^i(x)$。对于每项任务，本章使用线性模型作为评分函数，即有$\boldsymbol{f}^{(i)}(\boldsymbol{x}) = \boldsymbol{W}^{(i)^\top}\boldsymbol{x}$。

如引言所述，有必要考虑个性化评分的多样性，同时将多样性约束在合理范围内。实际上，可以通过由共识到个性的方式理解有限多样性：主流观点可由同一模式共享。不同人群可能存在不同的偏见或偏好，并由此产生偏离主流的观点。持相同偏好的人群形成小众群体，群体内部用户基于相似的物体特征子集形成相似的相对于主流观点的偏离。最终，群体中高度个性化的用户倾向于对群体观点持额外偏见。因此，提出对模型参数的层级化分解：$\boldsymbol{W}^{(i)} = \boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}$。其中，$\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$代表捕捉全局用户偏好的共识要素；$\boldsymbol{G}^{(i)} \in \mathbb{R}^{d\times 1}$为第$i$项任务的群体要素；$\boldsymbol{P}^{(i)}$为上述的用户特定要素。出于形式简洁，记$\boldsymbol{G} = [\boldsymbol{G}^{(1)}, \cdots, \boldsymbol{G}^{(U)}]$，且$\boldsymbol{G} \in \mathbb{R}^{d\times U}$。同理，有$\boldsymbol{P} = [\boldsymbol{P}^{(1)}, \cdots, \boldsymbol{P}^{(U)}]$，且$\boldsymbol{P} \in \mathbb{R}^{d\times U}$。参数分解的整体框架如图4.1所示。

综合上述设定，本章采用目标函数具有以下一般形式：

$$\min_{\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P}} \sum_{i=1}^{U} \ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)}) + \lambda_1\mathcal{R}_1(\boldsymbol{\theta}) + \lambda_2\mathcal{R}_2(\boldsymbol{G}) + \lambda_3\mathcal{R}_3(\boldsymbol{P}). \tag{4.1}$$

给定问题4.1，需进一步确定两个构成模块：

- 逐用户经验损失函数$\ell_i(\cdot,\cdot)$；
- 定义$\boldsymbol{W}$的先验约束，即正则化算子$\mathcal{R}_1(\boldsymbol{\theta})$、$\mathcal{R}_2(\boldsymbol{G})$和$\mathcal{R}_3(\boldsymbol{P})$

以下，本章将分别阐述上述两个模块。

4.2.3 正则化

对于共识要素$\boldsymbol{\theta}$，简单采用广泛使用的ℓ_2正则算子$\mathcal{R}_1(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$以降低模型复杂度。针对$\boldsymbol{G}$，本章追求用户特征协同聚类效应。(Xu 等, 2015)提出一种同时将矩阵$\boldsymbol{R}^{m\times n}$行和列聚类的方法：惩罚底部方阵$min\{n,m\}-\kappa$个奇异值之和。这启发本章采用以下形式的正则化算子：$\mathcal{R}_2(\boldsymbol{G}) = \sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})$。对于任一用户$i$，希望仅当其存在与共识、群体观点存在显著不一致时，$\boldsymbol{P}^{(i)}$存在非零列。因此，定义$\mathcal{R}_3(\boldsymbol{P}) = \|\boldsymbol{P}\|_{1,2}$以促进列的稀疏性。

4.2.4 经验损失及其计算方法

由于经验损失为逐用户方法，因此以下讨论聚焦于给定用户i，这并不影响结果的一般性。

4.2.4.1 经验损失

AUC的定义为随机抽样的阳性实例比随机抽样的阴性实例具有更高预测分数的概率。由于需最小化目标函数，以下考虑AUC的损失版本，即错排概率。即使数据分布未知，给定问题设定下的用户u_i和$\mathcal{S}_{+,i}, \mathcal{S}_{-,i}$，可得到AUC损失的经验估计：

$$\ell_{AUC}^{(i)} = \sum_{x_p\in\mathcal{S}_{+,i}} \sum_{x_q\in\mathcal{S}_{-,i}} \frac{I(\boldsymbol{x}_p,\boldsymbol{x}_q)}{n_{+,i}n_{-,i}},$$

其中离散错排惩罚$I(\boldsymbol{x}_p,\boldsymbol{x}_q)$具有以下形式：

$$I(\boldsymbol{x}_p,\boldsymbol{x}_q) = I_{[f^{(i)}(\boldsymbol{x}_p)>f^{(i)}(\boldsymbol{x}_q)]} + \frac{1}{2} I_{[f^{(i)}(\boldsymbol{x}_p)=f^{(i)}(\boldsymbol{x}_q)]}.$$

易知，$\ell_{AUC}^{(i)}$为给定数据集上用户i错排频率的精确值。然而，直接优化该指标为$\mathcal{NP}$难问题。因此，本文采用平方替代损失 $s(t) = (1-t)^2$ (Gao 等, 2016)。相应地，定义经验损失$\ell_i(\boldsymbol{f}^{(i)},\boldsymbol{y}^{(i)})$为

$$\ell_i(\boldsymbol{f}^{(i)},\boldsymbol{y}^{(i)}) = \sum_{x_p\in\mathcal{S}_{+,i}} \sum_{x_q\in\mathcal{S}_{-,i}} \frac{s\left(\boldsymbol{f}^{(i)}(\boldsymbol{x}_p) - \boldsymbol{f}^{(i)}(\boldsymbol{x}_q)\right)}{n_{+,i}n_{-,i}}.$$

4.2.4.2 高效AUC计算方法

直觉上，逐对AUC损失相较于逐例损失有着较高的计算复杂度。但是，重构ℓ_i可以有效降低由成对形式引起的计算负担。为此，首先定义图结构$\mathcal{G}^{(i)}$ =

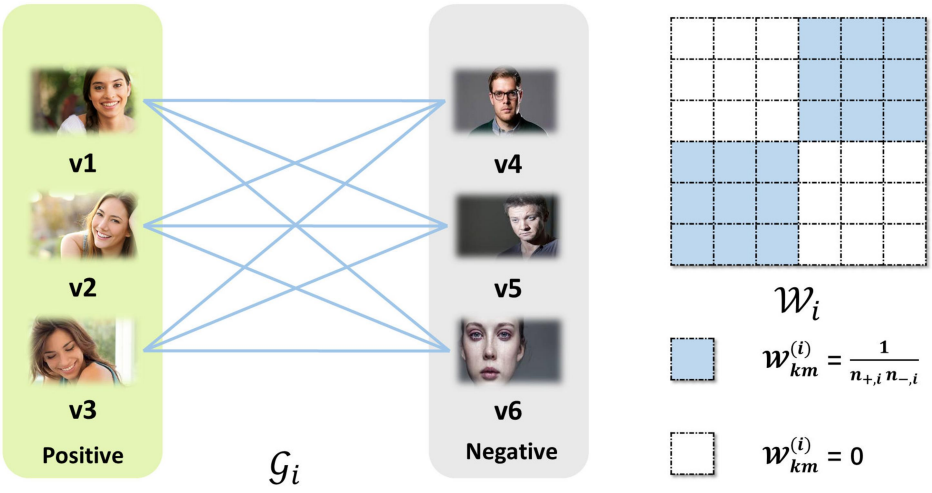


图 4.2 AUC计算图（以属性微笑的标注为例）

Figure 4.2 AUC computational graph (taking the annotation for attribute *smile* as an example)

$(\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{W}^{(i)})$。如图4.2所示，顶点集$\mathcal{V}^{(i)}$为$(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)})$中所有实例的集合；当且仅当$y_k^{(i)} \neq y_m^{(i)}$时，存在带有权重$\mathcal{W}_{km}^{(i)} = \frac{1}{n_{+,i}n_{-,i}}$的边$(k,m) \in \mathcal{E}^{(i)}$。给定$\mathcal{W}^{(i)}$，$\mathcal{G}_i$的拉普拉斯矩阵可表示为

$$\boldsymbol{L}^{(i)} = diag(\mathcal{W}^{(i)}\mathbf{1}) - \mathcal{W}^{(i)}.$$

由此，可重构经验损失为$\boldsymbol{L}^{(i)}$定义的二次型：

$$\ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)}) = \frac{1}{2}(\tilde{\boldsymbol{y}}^{(\boldsymbol{i})} - \boldsymbol{f}^{(i)})^\top \boldsymbol{L}^{(i)}(\tilde{\boldsymbol{y}}^{(\boldsymbol{i})} - \boldsymbol{f}^{(i)}),$$

其中$\tilde{\boldsymbol{y}}^{(\boldsymbol{i})} = \frac{\boldsymbol{y}^{(i)}+1}{2}$。进一步，以下命题提供加速计算$\boldsymbol{A}^\top \boldsymbol{L}^{(i)}\boldsymbol{B}$和$\boldsymbol{A}^\top \boldsymbol{L}^{(i)}$的方法：

命题 4.1. 给定$\boldsymbol{A} \in \mathbb{R}^{n\times a}$和$\boldsymbol{B} \in \mathbb{R}^{n\times b}$，其中$a, b$均为正整数，可在$\mathcal{O}\big(n_i(a+b+ab)\big) = \mathcal{O}(abn_i)$和$\mathcal{O}(an_i)$的时间复杂度内分别完成对$\boldsymbol{A}^\top \boldsymbol{L}^{(i)}\boldsymbol{B}$和$\boldsymbol{A}^\top \boldsymbol{L}^{(i)}$的计算。

注 4.1. 根据上述命题，可将$\boldsymbol{A}^\top \boldsymbol{L}^{(i)}\boldsymbol{B}$的复杂度自$O(abn_{+,i}n_{-,i})$降至$O(ab(n_{+,i} + n_{-,i}))$，同时可将$\boldsymbol{A}^\top \boldsymbol{L}$的复杂度自$O(an_{+,i}n_{-,i})$降至$O(a(n_{+,i} + n_{-,i}))$。

综上，最终的目标函数为

$$
\begin{aligned}
(P^*)\ \min_{\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P}}\ & \underbrace{\sum_i \sum_{\boldsymbol{x}_p\in\mathcal{S}_{+,i}} \sum_{\boldsymbol{x}_q\in\mathcal{S}_{-,i}} \frac{s\left(\boldsymbol{W}^{(i)\top}(\boldsymbol{x}_p-\boldsymbol{x}_q)\right)}{n_{+,i}n_{-,i}}}_{\mathcal{L}(\boldsymbol{W})} \\
& + \lambda_1 \underbrace{\|\boldsymbol{\theta}\|_2^2}_{\mathcal{R}_1(\boldsymbol{\theta})} + \lambda_2 \underbrace{\sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})}_{\mathcal{R}_2(\boldsymbol{G})} + \lambda_3 \underbrace{\|\boldsymbol{P}\|_{1,2}}_{\mathcal{R}_3(\boldsymbol{P})} \\
s.t\quad & \boldsymbol{W}^{(i)} = \boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}
\end{aligned}
\tag{4.2}
$$

为保证形式简洁，将经验损失表示为$\mathcal{L}(\boldsymbol{W})$，将$(P^*)$的目标函数表示为$\mathcal{F}(\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P})$。需注意$\mathcal{L}(\boldsymbol{W})$应为$\boldsymbol{\theta}$、$\boldsymbol{G}$和$\boldsymbol{P}$的函数。

4.3 模型优化

本章采用近端梯度法作为模型优化器。本节介绍该优化算法，并为$\mathcal{R}_2(\boldsymbol{G})$的近端算子提供闭式解。

对于每一迭代步骤k，考虑参照点$\boldsymbol{W}^{ref_k} = (\boldsymbol{\theta}^{ref_k}, \boldsymbol{G}^{ref_k}, \boldsymbol{P}^{ref_k})$，近端梯度法的变量更新可表示为

$$
\boldsymbol{\theta}^k := \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{2}\left\|\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}^k\right\|_2^2 + \frac{\lambda_1}{\rho_k}\|\boldsymbol{\theta}\|_2^2 \tag{4.3}
$$

$$
\boldsymbol{G}^k := \operatorname*{argmin}_{\boldsymbol{G}} \frac{1}{2}\left\|\boldsymbol{G}-\tilde{\boldsymbol{G}}^k\right\|_F^2 + \frac{\lambda_2}{\rho_k}\sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G}) \tag{4.4}
$$

$$
\boldsymbol{P}^k := \operatorname*{argmin}_{\boldsymbol{P}} \frac{1}{2}\left\|\boldsymbol{P}-\tilde{\boldsymbol{P}}^k\right\|_F^2 + \frac{\lambda_3}{\rho_k}\|\boldsymbol{P}\|_{1,2} \tag{4.5}
$$

其中$\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^{ref_k} - \frac{1}{\rho_k}\nabla_\theta\mathcal{L}(\boldsymbol{W}^{ref_k})$，$\tilde{\boldsymbol{G}}^k = \boldsymbol{G}^{ref_k} - \frac{1}{\rho_k}\nabla_{\boldsymbol{G}}\mathcal{L}(\boldsymbol{W}^{ref_k})$，$\tilde{\boldsymbol{P}}^k = \boldsymbol{P}^{ref_k} - \frac{1}{\rho_k}\nabla_P\mathcal{L}(\boldsymbol{W}^{ref_k})$；同时$\rho_k$由线性搜索策略选定，即保持更新$\rho_k = \alpha\rho_k, \alpha > 1$直至满足：

$$
\mathcal{L}(\boldsymbol{W}) < \mathcal{L}(\boldsymbol{W}^{ref_k}) + \Psi_{\rho_k}(D\boldsymbol{\theta}) + \Psi_{\rho_k}(D\boldsymbol{G}) + \Psi_{\rho_k}(D\boldsymbol{P}). \tag{4.6}
$$

其中$D\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}^{ref_k}$、$D\boldsymbol{G} = \boldsymbol{G} - \boldsymbol{G}^{ref_k}$、$D\boldsymbol{P} = \boldsymbol{P} - \boldsymbol{P}^{ref_k}$，且有

$$
\Psi_{\rho_k}(DA) = \left\langle \nabla_A\mathcal{L}(\boldsymbol{W}^{ref_k}), DA\right\rangle + \frac{\rho_k}{2}\left\langle DA, DA\right\rangle.
$$

注 4.2. $\nabla\mathcal{L}(\boldsymbol{W})$Lipschitz连续的性质保证了 ρ_k存在。需注意，选定参数最近的更新作为参照点，即有$\boldsymbol{W}^{ref_k} = \boldsymbol{W}^{k-1}$。

式(4.3)和式(4.5)的解可由ℓ_2范数和$\ell_{1,2}$范数的近端算子得到(Sra 等, 2012)。针对式(4.4)，以下给出闭式解。

由于$\sum_{\kappa+1}^{min\{d,U\}}\sigma_i(\boldsymbol{G})^2$非凸，传统方法使用交替迭代的方式优化(Xu 等, 2015)，效率较低且缺少理论保障。本章使用通用奇异值阈值框架(Lu 等, 2015; Lin 等, 2017b)，从而依据以下命题得到闭式最优解：

命题 4.2. 式 (4.4)的最优解为：

$$\boldsymbol{G}^* = \boldsymbol{U}\mathcal{T}_{\kappa,\frac{\lambda_3}{\rho_k}}(\boldsymbol{\Sigma})\boldsymbol{V}^\top, \tag{4.7}$$

其中$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$为$\tilde{\boldsymbol{G}}^k$的SVD分解，$\mathcal{T}_{\kappa,c}$将$\boldsymbol{\Sigma} = diag(\sigma_1,\cdots,\cdots,\sigma_{min\{d,U\}})$映射为对角矩阵，其大小与$\mathcal{T}_{\kappa,c}(\boldsymbol{\Sigma})_{ii} = (\frac{1}{2c+1})^{I[i>\kappa]}\sigma_i$相同。

4.4　理论分析

证明细节发布于Github主页[2]。

4.4.1　$\mathcal{L}(\boldsymbol{W})$梯度的Lipschitz连续性

上节指出，$\mathcal{L}(\boldsymbol{W})$梯度的Lipschitz连续性为线性搜索过程发现$\rho_k$的必要条件。以下定理表明，该性质为优化算法提供了理论保障。

定理 4.1 (Lipschitz连续梯度). 假设数据有界：

$$\forall i,\ \|\boldsymbol{X}^{(i)}\|_2 = \sigma_{X_i} < \infty,\ n_{+,i} \geq 1,\ n_{-,i} \geq 1.$$

给定任意两个不同的参数 $\boldsymbol{W}, \boldsymbol{W}'$，有：

$$\|\nabla\mathcal{L}(vec(\boldsymbol{W})) - \nabla\mathcal{L}(vec(\boldsymbol{W}'))\| \leq \gamma\Delta\boldsymbol{W} \tag{4.8}$$

其中，$\gamma = 3U\sqrt{(2U+1)}\max_i\left\{\frac{n_i\sigma_{X_i}^2}{n_{+,i}n_{-,i}}\right\}$，$vec(\boldsymbol{W}) = [\boldsymbol{\theta}, vec(\boldsymbol{G}), vec(\boldsymbol{P})]$，$\Delta\boldsymbol{W} = \|vec(\boldsymbol{W}) - vec(\boldsymbol{W}')\|$。

4.4.2　收敛性分析

由于正则化算子$\sum_{i=\kappa+1}^{min\{d,U\}}\sigma_i^2(\boldsymbol{G})$非凸，因此传统方法的次微分不完全适用于该问题。本章采用(Rockafellar 等, 2009; Liu 等, 2018)所提出的广义次微分。为保证次微分集非空，目标函数须满足下半连续性。而$\mathcal{L}(\boldsymbol{W})$、$\mathcal{R}_1(\boldsymbol{\theta})$和$\mathcal{R}_2(\boldsymbol{P})$均为

[2]https://joshuaas.github.io/publication

连续函数，故显然满足下半连续性。对于非凸算子$\sum_{i=\kappa+1}^{min\{d,U\}}\sigma_i^2(\boldsymbol{G})$，以下引理保证其连续性及下半连续性。

引理 4.1. 函数$\sum_{i=\kappa+1}^{min\{d,U\}}\sigma_i^2(\boldsymbol{G})$ 的连续性与$\boldsymbol{G}$一致。

接着，以下定理总结了所提出方法的收敛性质。此处定义$\Delta(\boldsymbol{\theta}^k)=\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k$、$\Delta(\boldsymbol{G}^k)=\boldsymbol{G}^{k+1}-\boldsymbol{G}^k$、$\Delta(\boldsymbol{P}^k)=\boldsymbol{P}^{k+1}-\boldsymbol{P}^k$。

定理 4.2. 假设初始解$\boldsymbol{\theta}^0,\boldsymbol{G}^0,\boldsymbol{P}^0$有界，则以下性质成立：

- 1）序列$\{\mathcal{F}(\boldsymbol{W}^k,\boldsymbol{U}^k)\}$非增，当以下条件满足时：$\forall k,\exists C_{k+1}>0$

$$\mathcal{F}(\boldsymbol{W}^{k+1},\boldsymbol{U}^{k+1})\leq\mathcal{F}(\boldsymbol{W}^k,\boldsymbol{U}^k)-C_{k+1}(\|\Delta(\boldsymbol{\theta}^k)\|_2^2+\|\Delta(\boldsymbol{G}^k)\|_F^2+\|\Delta(\boldsymbol{P}^k)\|_F^2) \tag{4.9}$$

- 2）$\lim_{k\to\infty}\boldsymbol{\theta}^k-\boldsymbol{\theta}^{k+1}=0,\ \lim_{k\to\infty}\boldsymbol{G}^k-\boldsymbol{G}^{k+1}=0,\lim_{k\to\infty}\boldsymbol{P}^k-\boldsymbol{P}^{k+1}=0$。
- 3）参数序列$\{\boldsymbol{\theta}^k\}_k,\{\boldsymbol{G}^k\}_k,\{\boldsymbol{P}^k\}_k$ 有界。
- 4）$\{\boldsymbol{\theta}^k,\boldsymbol{G}^k,\boldsymbol{P}^k\}_k$的极限点都是该问题的一个临界点。
- 5）$\forall\, T\geq 1,\exists\, C_T>0$:

$$\min_{0\leq k<T}\left(\|\Delta(\boldsymbol{\theta}^k)\|_2^2\right)\leq\frac{C_T}{T},\ \min_{0\leq k<T}\left(\|\Delta(\boldsymbol{G}^k)\|_F^2\right)\leq\frac{C_T}{T},\quad\min_{0\leq k<T}\left(\|\Delta(\boldsymbol{P}^k)\|_F^2\right)\leq\frac{C_T}{T}.$$

4.4.3 泛化界

定义参数集合Θ为：

$$\Theta=\left\{(\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P}):\sqrt{\mathcal{R}_1(\boldsymbol{\theta})}\leq\psi_1,\mathcal{R}_2(\boldsymbol{G})\leq\psi_2,\|\boldsymbol{G}\|_2\leq\sigma_{max}<\infty,\mathcal{R}_3(\boldsymbol{P})\leq\psi_3\right\} \tag{4.10}$$

则存在以下联合界。

定理 4.3. 假设$\exists\Delta_\chi>0$，且所有样本均满足$\|x\|\leq\Delta_\chi$。若定义

$$C=(\psi_1+\sqrt{\psi_2+\kappa\cdot\sigma_{max}^2}+\psi_3),$$

其中$\zeta=\Delta_\chi C$，则对于任意$\delta\in(0,1)$和任意$(\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P})\in\Theta$有以下结论成立：

$$\mathbb{E}_{\mathcal{D}}(\sum_i\ell_{AUC}^{(i)})\leq\mathcal{L}(\boldsymbol{W})+\sum_{i=1}^{U}\frac{B_1}{\sqrt{(n_i\chi_i(1-\chi_i))}}+B_2\sqrt{\frac{ln(\frac{2}{\delta})}{\sum_{i=1}^{U}n_i\chi_i(1-\chi_i)}} \tag{4.11}$$

以至少$1-\delta$概率成立，其中$B_1=8\sqrt{2}C\Delta_\chi(1+\zeta),B_2=10\sqrt{2}(1+\zeta)\zeta,\chi_i=\frac{n_{+,i}}{n_i}$。分布$\mathcal{D}=\otimes_{i=1}^{U}(\mathcal{D}_{+,i}\otimes\mathcal{D}_{-,i})$，其中对于用户$i$，$\mathcal{D}_{+,i}$和$\mathcal{D}_{-,i}$分别为正、负样本的条件分布。

注 4.3. 根据定理4.2，损失函数非增。对于所提出方法的解$(\boldsymbol{\theta}^*, \boldsymbol{G}^*, \boldsymbol{P}^*)$，有以下结论成立：

$$\sqrt{\mathcal{R}_1(\boldsymbol{\theta}^*)} \leq \sqrt{\frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda 1}}, \mathcal{R}_2(\boldsymbol{G}^*) \leq \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda 2} \mathcal{R}_3(\boldsymbol{P}^*) \leq \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda 3}$$

同时，由定理4.2可知$\boldsymbol{G}^*$有界。通过选定$\psi_1 = \sqrt{\frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda 1}}, \psi_2 = \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_2}, \psi_3 = \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda 3}$，本算法的解均属于$\Theta$。同时，上述所有解均以大概率取得较小的泛化误差，即期望离散AUC损失$\mathbb{E}_{\mathcal{D}}(\sum_i \ell_{AUC}^{(i)})$与在训练集$\mathcal{L}(W)$上的替代损失的差依速率$\mathcal{O}(\sum_{i=1}^{U} \frac{1}{\sqrt{(n_i \chi_i (1-\chi_i))}})$收敛。

4.5　实验

4.5.1　数据集

本章考虑以下数据集：

仿真数据集。考虑包含100名用户、500,000份标注结果的仿真数据集，其中样本特征和AUC评分均依据所提出方法的设定生成。对于每名用户，生成5000份样本得到矩阵$\boldsymbol{X}^{(i)} \in \mathbb{R}^{5000\times 80}$，且$\boldsymbol{x}_k^{(i)} \sim \mathcal{N}(0, \boldsymbol{I}_{80})$。为捕捉全局信息，设$\boldsymbol{\theta}$服从分布$\mathcal{U}(0,5) + \mathcal{N}(0, 0.5^2)$。为保证协同成簇性质，$\boldsymbol{G}$需满足协同成组结构。具体而言，为$\boldsymbol{G}$构造5个分块：$\boldsymbol{G}(1:20, 1:20)$、$\boldsymbol{G}(21:40, 21:40)$、$\boldsymbol{G}(41:50, 41:60)$、$\boldsymbol{G}(51:70, 61:80)$及$\boldsymbol{G}(71:80, 81:100)$。对于每一分块，元素依据分布$\mathcal{N}(C_i, 2.5^2)$逐元素采样生成，其中各簇质心$C_i \sim \mathcal{U}(0, 10)$由该簇内样本共享。同时，将不属于上述5个分块的元素设置为0。针对$\boldsymbol{P} \in \mathbb{R}^{d\times U}$，依据分布$\mathcal{U}(0, 10)$随机设定$\boldsymbol{P}(:, 1:5)$、$\boldsymbol{P}(:, 10:15)$、$\boldsymbol{P}(:, 20:25)$，并将剩余项设置为0。对于每位用户，通过$\boldsymbol{s}^{(i)} = \boldsymbol{X}^{(i)}(\boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}) + \epsilon^{(i)}$生成评分函数，其中$\epsilon^{(i)} \in \mathbb{R}^{5000\times 1}$、$\epsilon^{(i)} \sim \mathcal{N}(0, 0.01^2 \boldsymbol{I}_{5000})$。为生成每位用户$i$的标签$\boldsymbol{Y}^{(i)}$，选取评分前100的样例作为正样本（标签为1），其余作为负样本（标签为-1）。

Shoes数据集。数据集Shoes由(Kovashka 等, 2015)收集，其中包含了14,658份在线购物图片，对应7种属性，均由具有广泛兴趣和背景的用户标注得到。对于每类属性，至少分配了190名用户参与标注；而每位用户均标注了50幅图片，共计获得90,000份标注结果。为生成图片特征，将原始数据集提供的GIST和颜色直方图串联。为过滤冗余输入特征，在执行训练前通过主成分分析法（Principal Component Analysis, PCA），保留可解释99%样本方差的数据维度。同时，为消

除仅提供正标签的极端用户带来的影响，移除为某些类给出少于8份标注的用户。

Sun Attributes数据集。Sun Attributes (Patterson 等, 2012)数据集是著名的大规模场景属性数据集，其中包含约14,000张图片和102种可区分的属性。近期，(Kovashka 等, 2015)依靠数百名标注者收集了5种属性的个性化标注结果。每位标注者根据其个人理解与偏好标注了50幅图片，共计包括64,900份标注结果。数据集预处理与Shoes数据集完全相同；不同之处为将Inception-V3网络(Szegedy 等, 2016)倒数第二个全连接层的输出作为输入特征，而PCA分解仅保留可解释90%样本方差的数据维度

4.5.2 对比方法

本章考虑以下对比方法： **Robust Multi-Task Learning (RMTL)** (Chen 等, 2011)：RMTL旨在学习多任务中的无关任务，因此将模型参数分解为低秩、组稀疏结构。**Robust Multi-Task Feature Learning (rMTFL)** (Yu 等, 2007)：rMTFL假设模型可分解为两组：共享特征结构$\boldsymbol{P}$（$\ell_{1,2}$范数惩罚），和检测离群点的组系数结构$\boldsymbol{Q}$（施加于$\ell_{1,2}$转置的范数惩罚）。**Lasso**：带有ℓ_1范数正则化算子的多任务最小二乘法。**Joint Feature Learning (JFL)**(Nie 等, 2010)：JFL中所有模型共享同一组特征，因此通过$\ell_{1,2}$范数实现组稀疏约束。**The Clustered Multi-Task Learning Method (CMTL)** (Zhou 等, 2011)：CMTL将任务聚类为k个分组，并使用基于k-means的正则化算子实现该结构。**The task-feature coclusters based multi-task method (COMT)** (Xu 等, 2015)：COMT假设的特定于任务的组件具有特征-任务协同成簇结构。**Reduced Rank Multi-Stage multi-task learning (RAMU)** (Han 等, 2016)：RAMU采用capped 迹范数正则化仅最小化小于自适应阈值的奇异值。

4.5.3 实验细节

所有实验均依据训练集和验证集（共占实例总数的85％）调参，并记录测试集上15轮重复实验的结果。

表 4.1 仿真数据集上AUC性能对比

Table 4.1 AUC Comparison on Simulation Dataset

算法	RMTL	rMTFL	LASSO	JFL
均值	83.48	83.45	83.57	83.49
算法	CMTL	COMT	RAMU	Ours
均值	83.47	83.44	83.50	**99.65**

表 4.2 程序运行时间比较

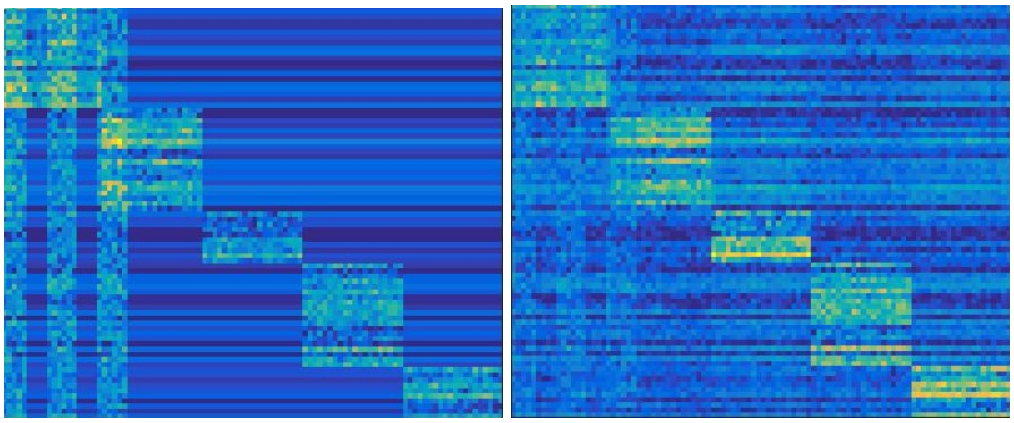
Table 4.2 Running Time Comparison [3]

ratio	20%	40%	60%	80%	100%
Orginal	18.57	74.22	151.86	268.55	nan
Ours	**3.06**	**5.50**	**8.65**	**12.46**	**15.82**

表 4.3 基于AUC的性能对比

Table 4.3 Performance Comparison based on the AUC metric

Alg	Attibutes											
	Shoes							Sun				
	BR	CM	FA	FM	OP	ON	PT	CL	MO	OP	RU	SO
RMTL	79.31	84.99	66.90	85.08	75.67	67.22	75.14	69.36	62.71	75.28	67.91	69.23
rMTFL	70.90	83.78	67.27	85.91	73.71	65.21	77.11	69.27	62.15	75.80	68.16	68.76
LASSO	68.46	80.48	65.90	84.01	71.47	64.60	75.08	67.64	61.83	75.39	68.57	69.13
JFL	72.00	83.10	67.26	85.93	73.02	65.39	77.09	68.63	61.94	75.00	67.17	68.78
CMTL	74.54	85.16	68.21	85.32	75.06	68.17	77.62	72.55	66.61	79.78	72.34	72.82
COMT	84.24	88.68	69.66	89.19	80.93	72.99	80.62	70.69	63.72	76.93	69.43	70.44
RAMU	78.33	84.58	65.78	84.68	75.25	66.72	73.50	72.95	69.25	79.81	74.39	72.50
Ours	**92.95**	**90.92**	**73.24**	**92.65**	**87.95**	**81.07**	**86.22**	**79.31**	**78.19**	**86.50**	**81.88**	**78.98**



(a) Ground-Truth **(b) Learned Parameter**

图 4.3 所提方法恢复预期参数结构的能力

Figure 4.3 The Potential of our proposed method to Recover the Expected Parameter Structure

4.5.4 实验结果

仿真数据集： 表4.1记录了仿真数据集上所有相关算法的性能。结果显示，本章提出的算法优于所有对比方法。具体而言，所提出方法的AUC达到99.65%，而第二名算法仅达到83.57%。

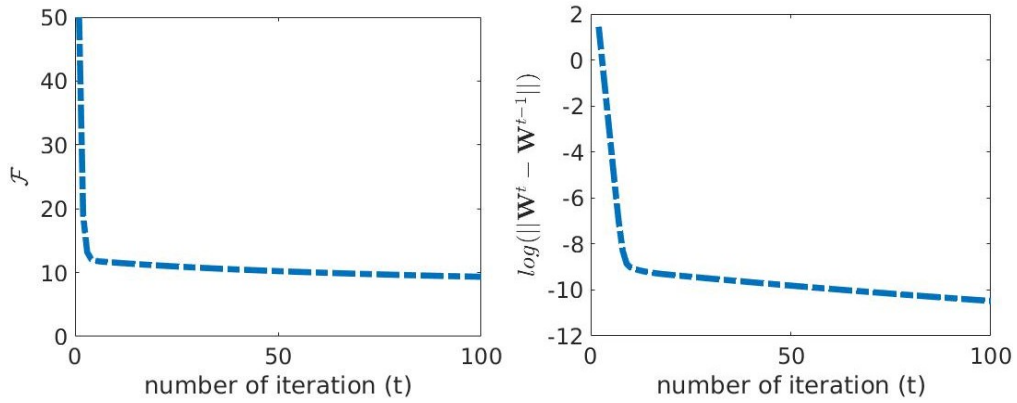
除了泛化性能，同时验证了所提出算法恢复参数$\boldsymbol{W}$预期结构的能力。在相同的仿真数据集上，通过比较算法习得参数与真实参数（如图4.3所示），可以发现所提出算法能够大致恢复期望的分组结构。

定理4.2证明了所提出方法的收敛过程。为验证该理论发现，图4.4展示了损失和参数随着训练轮数的演化过程。如图4.4-(a)所示，损失随着训练过程的推进逐渐下降；而在图4.4-(b)中，参数差$log(\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^{t}\|)$同样逐渐下降。上述所有观察均与理论结果相吻合。

为验证提出的AUC计算方法的效率，需记录使用和不使用加速算法时的程序运行时间。表4.2记录了使用不同比率的数据集作为训练集时的运行时间。可以发现，当训练样本数量增加时，不使用加速算法导致运行时间迅速提高。当使用全体数据集训练时，受限于内存（24GB），程序无法在1小时内结束（用nan代替实际运行时间）。相反，所提出的加速方法取得了20倍加速。

Shoes数据集： 表4.3左半部分展示了15次重复实验的平均结果（BR：棕色，

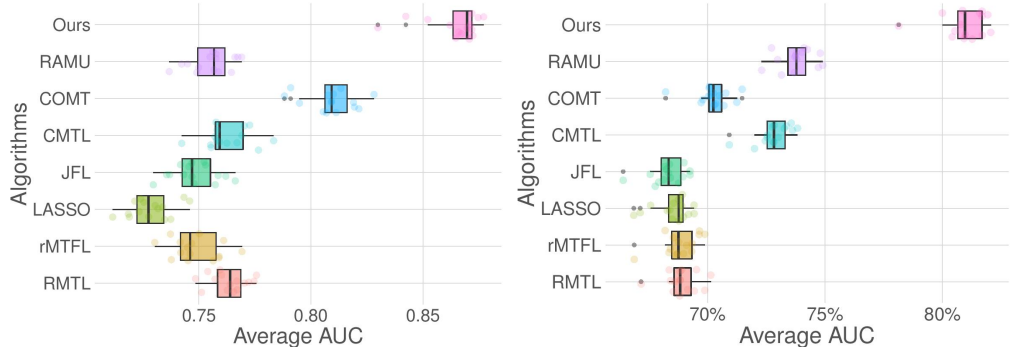
[3]Original代表原始的AUC计算时间，而Ours代表加速后时间（单位：秒）。



(a) Loss Convergence　　**(b) Paramter Convergence**

图 4.4 仿真数据集上的算法收敛曲线：a）损失收敛曲线，而b）参数收敛曲线

Figure 4.4 The Convergence Curve On Simulation Dataset: a) loss convergence curve, whereas b) parameters convergence curve



(a) Shoes Dataset　　**(b) Sun Attribute Dataset**

图 4.5 真实数据集上所有属性的平均性能

Figure 4.5 Average performances on all attributes of two real-world datasets

CM：舒适的，FA：时尚的，FM：正式的，OP：开放的，ON：华丽的，PT：尖的）。图4.5通过箱图展示了7类属性重复15次的实验结果。因此可以得出结论，所提出的算法在性能上优于所有基准算法。

Sun Attributes数据集： 表4.3右半部分展示了15次重复实验的平均结果（CL：杂乱无章的，MO：现代的，OP：开放领域，RU：乡村的，SO：平静的）。图4.5-(b)通过箱图展示了5类属性重复15次的实验结果。与Shoes数据集相似，所提出的算法优于所有基准方法。

4.6 小结

本章提出一种多任务模型用以学习特定于用户的属性理解。该模型一方面通过多层次参数分解建模主流共识、群体观点与个人意见，另一方面设计基于AUC的损失函数学习复杂的属性偏好。进一步，本章提出一种高效的AUC计算方法，从而显著降低损失和梯度的计算复杂度。最终，理论与实验结果均验证了本章所提出方法的有效性。

第5章　基于任务-特征协同学习的多任务AUC优化方法及应用

5.1　引言

现有机器学习方法通常依赖大量训练数据。然而实际应用中的训练样本往往难以收集，如何利用小规模数据集提高模型性能因此成为亟待解决的问题。面临多个相关任务时，一个有效的解决方案是多任务协同学习（Multi-Task Learning, MTL）范式。一般而言，多任务协同学习范式通过构造一定的约束条件促进全部/部分任务间的知识迁移，并完成多个任务的联合训练。早期研究工作(Heskes, 1998)表明，当单个任务的标注不足时，联合训练多个相关任务可以显著提升模型泛化性能。现如今，多任务协同学习范式已广泛应用于机器学习领域，成为包括场景解析(Xu 等, 2018)、属性学习(Cao 等, 2018)、文本分类(Liu 等, 2017a)、序列标记(Lin 等, 2018)、旅行时间(Li 等, 2018)估计等众多应用场景的重要组成部分。

MTL的基本观点认为，在多个任务间进行知识共享能够提升模型的泛化性能。基于该观点，已有大量研究工作探索如何促进不同任务间的有效知识共享。早期研究在所有任务上共享知识，例如，(Argyriou 等, 2008a)通过构建通用的、稀疏的特征促使不同任务之间进行知识迁移。然而，(Kang 等, 2011)指出，当存在不相关任务时，与不相关任务或难任务共享通用的特征常常导致模型性能下降，这种现象被称作*负迁移*。为解决*负迁移*问题，近期研究工作主要从以下两个方向着手。

第一个方向首先将不同任务进行分组，并将其视为一个聚类问题。作为该方向的早期工作，(Thrun 等, 1996)首先构建任务簇，并分别学习隶属于各个簇的模型参数。鉴于该分阶段式方法无法保证簇和模型参数的最优性，许多研究进而探索如何将聚类与多任务学习纳入统一框架中。总体而言，该方向工作可分为两大类。第一类方法采用贝叶斯学习框架，假设任务特异的参数服从聚类假设，例如高斯先验(Bakker 等, 2003)和Dirichlet过程先验(Xue 等, 2007; Qi 等, 2008; Ni 等, 2007)的混合。第二类方法将聚类问题形式化为正则化算子。具体而言，构建正则化算子以：(a)惩罚小簇间方差和大簇内方差，(b)松弛整数规划问题，(c)鼓励结构稀疏性(Han 等, 2015; Zhou 等, 2016; McDonald 等, 2016)， (d) 构

建了相应的正则化算子(Kumar 等, 2012; Maurer 等, 2013)。

另一方向认为知识迁移是非对称的。事实上，从简单任务到困难任务的知识迁移一般较为安全，而从困难任务到简单任务的知识迁移则是负迁移的主要来源。受此启发，(Lee 等, 2016)首次提出在MTL方法中考虑非对称性。该工作假定任务参数位于由其自身张成的列空间中，然后利用每个任务的稀疏表示系数实现非对称性。随后，一些工作致力于从以下几个方面对该框架进行改进：(a)自适应于预测器相关系数的惩罚项(Wang 等, 2015)；(b) 潜在任务表示学习(Liu 等, 2017b; Grave 等, 2011)； (c) 分组约束(Oliveira 等, 2019)；(d) 鲁棒约束(Yao 等, 2019)。

大多数现有方法仅将负迁移问题建模为任务分组问题。然而，即使确保对任务进行合理分组，在不同任务组间共享冗余特征仍存在过拟合风险。更为具体地，负迁移可能同时发生在不同任务之间和不同特征之间。受此启发，本章旨在通过对特征进行分组，从而实现更加通用的负迁移抑制解决方案。为此，本章在MTL中引入任务和特征的协同分割，提出一种任务特征协同学习（Task-Feature Collaborative Learning, TFCL）框架。具体来说，通过三个步骤构建该框架。

Step 1 提出一个基本模型，构建以特征和任务为节点的二部图并提出新的正则化算子约束邻接矩阵的块对角结构，从而实现联合分组的目标。

Step 2 鉴于基本模型对应的优化问题(记作($\boldsymbol{P}$))是非凸非光滑的，提出一个替代问题(记作($\boldsymbol{P}^\star$))并构建优化算法进行求解。证明在一定假设条件下，问题 ($\boldsymbol{P}$) 和问题($\boldsymbol{P}^\star$)能够同时达到全局收敛。除收敛分析外，优化算法产生的中间解隐式地提供了每个特征和任务的嵌入表达，基于这些嵌入表达，进一步表明，通过选择合适的参数，优化算法可以保证预期的块对角结构。实现了跨特征、跨任务的分组效果，从而抑制跨组间的负迁移。

Step 3 进一步聚焦于个性化属性预测的具体问题。个性化属性预测任务将单一给定用户的个性化属性预测视为一个任务。为获得更加灵活的模型，同时考虑：(a)捕获用户共享兴趣的共识因子，该因子允许分组中存在跨组重叠，(b) 基本模型中的联合分组因子， (c) 在分组中排除异常用户/任务的异常因子。在此基础上，证明该模型继承了TFCL的所有理论特性。

本章的主要贡献包括以下三个方面：

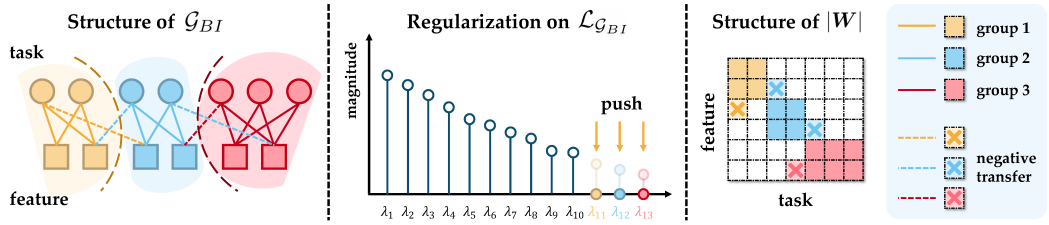


图 5.1 任务-特征协同学习框架的基本模型示意图

Figure 5.1 Illustration of the Base Model of the TFCL Framework.

(一) 提出任务特征协同学习框架，核心组件为一个基于任务-特征协同分割的新块对角正则化算子，从而能够同时在任务和特征级别上探索更一般的负迁移效应。

(二) 设计一个在全局收敛性上具有理论保障的优化算法，同时给出实现恢复块对角结构的理论保障。

(三) 最后，针对个性化属性预测任务，基于基本TFCL模型提出更实用、更灵活的扩展模型。

5.2　相关工作

本节简要回顾块对角结构学习、多任务学习两个方面有关工作的最新进展。

5.2.1　块对角结构学习

块对角结构学习的有关研究可追溯到聚类学习。聚类学习旨在以无监督的方式将不同数据点划分到不同的簇中。作为典型的聚类方法，基于图的聚类方法(例如，谱聚类 (Bach 等, 2004; Ng 等, 2002) 和子空间聚类 (Elhamifar 等, 2009; Liu 等, 2013)) 通常包含两个阶段：(a) 首先得到一个样本-样本邻接矩阵以捕捉不同样本之间的相关性。(b) 给定邻接矩阵，将聚类问题形式化为一个图分割问题并最小化归一化割(normalized cut)的谱松弛。基于该框架，当邻接矩阵具有明显的块对角结构时，每个对角块就代表一个簇。因此，利用邻接矩阵的块对角结构可以显著提高此类基于图的聚类方法的性能。受此启发，研究人员开始探索能够隐式保留邻接矩阵的块对角结构的正则化算子(Li 等, 2015a; You 等, 2016; Xin 等, 2017; Wipf 等, 2016; Wang 等, 2011; Favaro 等, 2011; Lu 等, 2018)。然而，正如(Lu 等, 2019)所述，隐式正则化算子无法处理来自输入特征零空间的非对角

线噪声。随后，(Nie 等, 2014, 2016)首次提出在基于图的聚类框架中开发显式块对角正则化算子以更好地解决该问题。最近，(Lu 等, 2019; Xie 等, 2017; Yang 等, 2019b)将其引入到子空间聚类框架中以构建自表达层。在该研究方向中，与本章最相关的是显式正则化算子的有关工作。然而，两者的不同之处在于以下两个方面。首先，既往工作针对同构样本，块对角属性仅适用于第i行/列均代表相同类型样本的方阵，而本章所提出的任务-特征联合分组适用范围更广。具体而言，本章提出一个广义块对角结构学习框架，可适用于任意大小的、第i行/列代表不同类型顶点的矩阵。其次，优化方法方面，既往工作仅提供一个子序列收敛保证，尚未有全局收敛性质。相反，通过对替代问题进行重构，本章所提出的方法具有全局收敛性保障。小节5.4.3更详细地讨论了有关工作与本章方法之间的联系。

5.2.2 多任务学习

引言一节简要回顾了针对负迁移问题的有关工作，本小节进一步讨论多任务学习方法有关方法。首先，从结构学习的角度来看，虽然引言所述非对称多任务学习方法(Lee 等, 2016; Liu 等, 2017b; Yao 等, 2019; Oliveira 等, 2019) 也利用了块对角结构，然而，正如第5.2.1节所述，该类方法仅适用于同构块对角结构，因此不适用于同时考虑任务结点与特征结点的异构块对角结构。从任务-特征协同学习的角度来看，在已知范围内，目前仅有两个多任务学习方法对任务-特征的联合分组结构进行显式建模。(Zhong 等, 2012)首先考虑不同特征在不同任务中的作用；然而，该方法分别处理各个特征，并未考虑特征之间的复杂关系。(Xu 等, 2015) 转而基于一个协同聚类诱导的正则化算子学习任务-特征的相关性。然而，该正则化算子缺乏能够显式恢复块对角结构的理论保障，且未能显式地建模任务-特征协同聚类与负迁移之间的联系。

5.3 框架介绍

本节主要介绍所提出的任务-特征协同学习框架的基本模型，其总体结构如图5.1所示。本节假设能够将任务和特征同时划分到不同组中；第5.4.4小节不再受限于该假设，并纳入异常任务和共识特征对TFCL模型进行扩展。

本章采用的符号及其说明如下。$\mathbb{S}_m$表示所有属于$\mathbb{R}^{m\times m}$的实对称矩阵集合。给定任意矩阵$\boldsymbol{A}\in\mathbb{S}_N$,将$\boldsymbol{A}$的$N$个特征根以降序排列，即，$\lambda_1(\boldsymbol{A})\le\lambda_2(\boldsymbol{A})\le\cdots\le\lambda_N(\boldsymbol{A})$。$\langle\cdot,\cdot\rangle$ 表示向量或矩阵内积。给定两个矩阵$\boldsymbol{A}$和$\boldsymbol{B}$，$\boldsymbol{A}\oplus\boldsymbol{B}$表示两个矩阵的直和。若$\boldsymbol{A}-\boldsymbol{B}$为半正定矩阵，记$\boldsymbol{A}\succeq\boldsymbol{B}$。 $\mathbb{U}(a,b)$ 代表均匀分布，$\mathbb{N}(\mu,\sigma^2)$代表正态分布。对于给定矩阵$\boldsymbol{A}\in\mathbb{R}^{m\times n}$，定义零空间$\mathcal{N}(\boldsymbol{A})=\{\boldsymbol{x}\in\mathbb{R}^n:\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}\}$。给定任意矩阵$\boldsymbol{A}\in\mathbb{S}_m$，若$\lambda$是其一个特征根，记$\mathbb{EIG}_A(\lambda)=\mathcal{N}(\boldsymbol{A}-\lambda\boldsymbol{I})$ 为特征根λ对应的特征向量张成的子空间。给定矩阵$\boldsymbol{A}=[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_n]$，记$\boldsymbol{A}_{m:n}=[\boldsymbol{a}_m,\cdots,\boldsymbol{a}_n]$。给定集合$\mathcal{A}$，若$x\in\mathcal{A}$，记作$\iota_{\mathcal{A}}(x)=0$；反之，$\iota_{\mathcal{A}}(x)=+\infty$。

假设有T个待学习任务，给定训练集:

$$\mathcal{S}=\left\{(\boldsymbol{X}^{(1)},\boldsymbol{y}^{(1)}),\cdots,(\boldsymbol{X}^{(T)},\boldsymbol{y}^{(T)})\right\}.$$

对于集合$\mathcal{S}$，$\boldsymbol{X}^{(i)}\in\mathbb{R}^{n_i\times d}$ 为第i个任务的输入特征矩阵，其中n_i表示第i个任务包含的实例数，d表示特征维度；矩阵$\boldsymbol{X}^{(i)}$的每一行代表一个实例的特征向量；$\boldsymbol{y}^{(i)}$表示第i个任务对应的响应或输出；给定任务i，建立线性模型$\boldsymbol{g}^{(i)}(x)=\boldsymbol{W}^{(i)\top}\boldsymbol{x}$，并记模型参数矩阵$\boldsymbol{W}=[\boldsymbol{W}^{(1)},\cdots,\boldsymbol{W}^{(T)}]\in\mathbb{R}^{d\times T}$；在标准的多任务学习范式中，$\boldsymbol{W}$ 可通过如下正则化约束问题求得：

$$\underset{\boldsymbol{W}}{\operatorname{argmin}}\,\mathcal{J}(\boldsymbol{W})+\alpha\cdot\Omega(\boldsymbol{W}),$$

其中$\mathcal{J}(\boldsymbol{W})=\sum_i\ell_i$，且$\ell_i$ 表示第i个任务的经验风险，$\Omega(\boldsymbol{W})$为正则项。以下推导合适的正则化算子，进而从任务和特征两个层面抑制负迁移。

基于线性模型，对于第i个任务，可通过$\hat{y}=\boldsymbol{W}^{(i)\top}\boldsymbol{x}=\sum_{j=1}^{d}W_{ij}x_j$预测其响应。若$W_{ij}=0$，则$\hat{y}$与第$j$个特征无关；若$|W_{ij}|$越大，则表示$\hat{y}$ 越依赖于第j个特征。可见，$|W_{ij}|$可直接特征i和任务j的相关性。为抑制跨任务、跨特征间的负迁移，TFCL将任务和特征自动划分为多个分组，其中每个组仅包含相关的任务和特征。若特征i和任务j不属于同一组且$W_{ij}\neq0$，则代表发生了负迁移。受此启发，本章旨在将任务和特征同时划分为k个组，从而当特征i和任务j不在同一组时，$|W_{ij}|$接近于0。为此，以下借助辅助二部图构建实现上述约束的正则化项算子。

具体而言，定义二部图$\mathcal{G}_{BI}=(\mathcal{V}_{BI},\mathcal{E}_{BI},\boldsymbol{A}_{BI})$，其中$\mathcal{V}_{BI}=\mathcal{V}_T\cup\mathcal{V}_F$表示顶点

集，$\mathcal{V}_T$和$\mathcal{V}_F$分别表示任务和特征集合；$\mathcal{G}_{BI}$的邻接矩阵定义为：

$$\boldsymbol{A}_{BI} = \begin{bmatrix} \boldsymbol{0} & |\boldsymbol{W}| \\ |\boldsymbol{W}|^\top & \boldsymbol{0} \end{bmatrix};$$

定义边集为$\mathcal{E}_{BI} = \{(i,j)|\boldsymbol{A}_{BI_{i,j}} > 0\}$。引入拉普拉斯矩阵

$$\mathcal{L}_{\mathcal{G}_{BI}} = diag(\boldsymbol{A}_{BI}\boldsymbol{1}) - \boldsymbol{A}_{BI}.$$

给定$\mathcal{G}_{BI}$，以下基于谱图论推导满足要求的正则化算子。根据谱图论，为保证将任务与特征顶点划分为k组，只需保证二部图 $\mathcal{G}_{BI}$存在k个连通分量。下述定理说明，保证二部图有k个连通分量等价于约束拉普拉斯矩阵$\mathcal{L}_{\mathcal{G}_{BI}}$零空间的维度为$k$。

定理 5.1. (Argyriou 等, 2008b) $\dim(\mathcal{N}(\mathcal{L}_{\mathcal{G}_{BI}})) = k$，即拉普拉斯矩阵$\mathcal{L}_{\mathcal{G}_{BI}}$的零特征根的重数为$k$，当且仅当二部图$\mathcal{G}_{BI}$包含$k$个连通分量。进一步，记$\mathcal{G}(i)$为隶属于第$i$个连通分量的任务和特征顶点集，则有：

$$\mathbb{EIG}_{\mathcal{L}_{\mathcal{G}_{BI}}}(0) = span(\boldsymbol{\iota}_{\mathcal{G}(1)}, \boldsymbol{\iota}_{\mathcal{G}(2)}, \cdots, \boldsymbol{\iota}_{\mathcal{G}(N)}),$$

其中 $\boldsymbol{\iota}_{\mathcal{G}(i)} \in \mathbb{R}^{(d+T)\times 1}$；当 $j \in \mathcal{G}(i)$时，$[\boldsymbol{\iota}_{\mathcal{G}(i)}]_j = 1$，反之$[\boldsymbol{\iota}_{\mathcal{G}(i)}]_j = 0$。

上述定理给出了实现本章目标的方法：构建正则化算子使得拉普拉斯矩阵的最小k个特征根为0。令$N = d + T$ 为二部图$\mathcal{G}_{BI}$的顶点数，该正则化算子等价于约束$rank(\mathcal{L}_{\mathcal{G}_{BI}}) = N - k$。然而直接优化该约束为NP难问题，因此本节将之松弛为最小化最小k个特征根之和，即$\sum_{i=1}^{k}\lambda_i(\mathcal{L}_{\mathcal{G}_{BI}})$。根据特征根的变分性质(Fan, 1949)，最小化$\mathcal{L}_{\mathcal{G}_{BI}}$ 的最小k个特征根之和可形式化为以下问题：

$$\sum_{i=1}^{k}\lambda_i(\mathcal{L}_{\mathcal{G}_{BI}}) = \min_{\boldsymbol{E}} tr(\boldsymbol{E}\mathcal{L}_{\mathcal{G}_{BI}}\boldsymbol{E}^\top), s.t.\ \boldsymbol{E}^\top\boldsymbol{E} = \boldsymbol{I}_k. \tag{5.1}$$

由于非凸约束$\boldsymbol{E}^\top\boldsymbol{E} = \boldsymbol{I}_k$的存在，上述问题是非凸的。最近有研究提出了该问题的一个等价凸形式：

定理 5.2. 令 $\Gamma = \{\boldsymbol{U} : \boldsymbol{U} \in \mathbb{S}_N,\ \boldsymbol{I} \succeq \boldsymbol{U} \succeq \boldsymbol{0}, tr(\boldsymbol{U}) = k\}$，则对于$\forall \boldsymbol{A} \in \mathbb{S}_N$:

$$\sum_{i=1}^{k}\lambda_i(\boldsymbol{A}) = \min_{\boldsymbol{U}\in\Gamma}\langle \boldsymbol{A},\ \boldsymbol{U}\rangle, \tag{5.2}$$

且最小值在$\boldsymbol{U} = \boldsymbol{V}_k\boldsymbol{V}_k^\top$处取到，其中$\boldsymbol{V}_k$代表矩阵$\boldsymbol{A}$的最小$k$个特征根对应的特征向量。

证明. 记矩阵$\boldsymbol{A}$ 的特征根分解为：

$$\boldsymbol{A} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^{\top},\ \Lambda = diag(\lambda_1(\boldsymbol{A}), \cdots, \lambda_N(\boldsymbol{A})). \tag{5.3}$$

对于可行集Γ中的任意一个$\boldsymbol{U}$，有以下式子成立: $\langle \boldsymbol{A}, \boldsymbol{U}\rangle = \sum_i C_{ii}\lambda_i(\boldsymbol{A})$, 其中$\boldsymbol{C} = \boldsymbol{Q}^{\top}\boldsymbol{U}\boldsymbol{Q}$。鉴于矩阵$\boldsymbol{C}$ 和矩阵$\boldsymbol{U}$有相同的特征根，因此，当且仅当$\boldsymbol{U} \in \Gamma$时，$\boldsymbol{C} \in \Gamma$成立。进而下式成立：

$$\min_{\boldsymbol{U}\in\Gamma} \langle \boldsymbol{A}, \boldsymbol{U}\rangle \Longleftrightarrow \min_{\boldsymbol{C}\in\Gamma} \sum_i C_{ii}\lambda_i(\boldsymbol{A}). \tag{5.4}$$

定义$\boldsymbol{e}^i \in \mathbb{R}^{N\times 1}$，$\boldsymbol{e}^i_i = 1$ 且对于任意$s \neq i$，$\boldsymbol{e}^i_s = 0$，可得：

$$C_{ii} = \frac{\boldsymbol{e}^{i\top}\boldsymbol{C}\boldsymbol{e}^i}{\boldsymbol{e}^{i\top}\boldsymbol{e}^i}.$$

于是，根据矩阵$\boldsymbol{C}$的特征根极值属性，可得如下不等式：

$$\begin{aligned} 0 \leq \lambda_1(\boldsymbol{C}) = \min_{\boldsymbol{x}} \frac{\boldsymbol{x}^{\top}\boldsymbol{C}\boldsymbol{x}}{\boldsymbol{x}^{\top}\boldsymbol{x}} \leq C_{ii} \\ \leq \max_{\boldsymbol{x}} \frac{\boldsymbol{x}^{\top}\boldsymbol{C}\boldsymbol{x}}{\boldsymbol{x}^{\top}\boldsymbol{x}} = \lambda_N(\boldsymbol{C}) \leq 1. \end{aligned} \tag{5.5}$$

当$C_{ii} = 0(i > k)$，$C_{ii} = 1(i \leq k)$时，公式 (5.4) 取到最小值$\sum_{i=1}^{k} \lambda_i(\boldsymbol{A})$。从而可证$\sum_{i=1}^{k} \lambda_i(\boldsymbol{A}) = \min_{\boldsymbol{U}\in\Gamma} \langle \boldsymbol{A}, \boldsymbol{U}\rangle$。

以下只需通过 $\sum_{i=1}^{k} \lambda_i(\boldsymbol{A}) = \langle \boldsymbol{A}, \boldsymbol{U}\rangle$证明 $\boldsymbol{U} = \boldsymbol{V}_k\boldsymbol{V}_k^{\top}$ 为最优解。由于$\boldsymbol{V}_k$ 为矩阵$\boldsymbol{A}$的最小k个特征根对应的特征向量，因此$\boldsymbol{Q} = [\boldsymbol{V}_k^{\perp}, \boldsymbol{V}_k]$，其中 $\boldsymbol{V}_k^{\perp}$ 表示最大$N-k$个特征根对应的特征向量，且满足 $\boldsymbol{V}_k^{\top}\boldsymbol{V}_k^{\perp} = \boldsymbol{0}$，$\boldsymbol{V}_k^{\perp\top}\boldsymbol{V}_k = \boldsymbol{0}$。由此可得：

$$\begin{aligned} \boldsymbol{C} = \boldsymbol{Q}^{\top}\boldsymbol{U}\boldsymbol{Q} &= \begin{bmatrix} \boldsymbol{V}_k^{\top} \\ \boldsymbol{V}_k^{\perp\top} \end{bmatrix} \boldsymbol{V}_k\boldsymbol{V}_k^{\top}[\boldsymbol{V}_k, \boldsymbol{V}_k^{\perp}] \\ &= \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \end{aligned} \tag{5.6}$$

后续证明过程遵循$\sum_i C_{ii}\lambda_i(\boldsymbol{A}) = \sum_{i=1}^{k} \lambda_i(\boldsymbol{A})$的证明，此处不再赘述。

□

结合经验风险$\mathcal{J}(\boldsymbol{W})$、正则化项和模型参数$\boldsymbol{W}$的惩罚项$\ell_2$，给出本章优化问题 $(\boldsymbol{P})$的形式化表述如下：

$$(\boldsymbol{P})\ \min_{\boldsymbol{W},\boldsymbol{U}\in\Gamma}\ \mathcal{J}(\boldsymbol{W}) + \alpha_1 \cdot \left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U} \right\rangle + \frac{\alpha_2}{2} \cdot \|\boldsymbol{W}\|_F^2. \tag{5.7}$$

5.4 模型优化

由于原问题($\boldsymbol{P}$)不易求解，本节构造其替代问题($\boldsymbol{P}^{\star}$)，表为：

$$(\boldsymbol{P}^{\star})\ \min_{\boldsymbol{W},\boldsymbol{U}\in\Gamma}\left\{\begin{array}{l}\mathcal{J}(\boldsymbol{W})+\alpha_1\cdot\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle+\dfrac{\alpha_2}{2}\cdot\|\boldsymbol{W}\|_F^2\\+\dfrac{\alpha_3}{2}\cdot\|\boldsymbol{U}\|_F^2\end{array}\right\}. \tag{5.8}$$

以下说明，在一定条件下，优化替代问题足以保证取得原问题($\boldsymbol{P}$)的一个临界点。

由于$\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle$关于参数$\boldsymbol{W}$非凸非光滑，因此采用近端梯度下降法（Proximal Gradient Decent, PGD）(Beck 等, 2009)进行求解。首先，假设损失函数的梯度$\nabla_{\boldsymbol{W}}\mathcal{J}(\boldsymbol{W})$是$\varrho$-Lipschitz连续的。根据近端梯度下降法，对于第t步迭代，给定常数$C>\varrho$和上一步的解$\boldsymbol{W}^{t-1}$，可通过求解如下问题更新参数$\boldsymbol{W}^t$和$\boldsymbol{U}^t$：

$$(\boldsymbol{Prox})\ \min_{\boldsymbol{W},\boldsymbol{U}\in\Gamma}\left\{\begin{array}{l}\dfrac{1}{2}\left\|\boldsymbol{W}-\widetilde{\boldsymbol{W}}^t\right\|_F^2+\dfrac{\alpha_1}{C}\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle\\+\dfrac{\alpha_2}{2C}\cdot\|\boldsymbol{W}\|_F^2+\dfrac{\alpha_3}{2C}\|\boldsymbol{U}\|_F^2\end{array}\right\}, \tag{5.9}$$

其中，$\widetilde{\boldsymbol{W}}^t=\boldsymbol{W}^{t-1}-\dfrac{1}{C}\nabla_{\boldsymbol{W}}\mathcal{J}(\boldsymbol{W}^{t-1})$。

5.4.1 子问题求解

求解$(\boldsymbol{Prox})$包括两个子问题：（1）给定参数$\boldsymbol{W}$，优化$\boldsymbol{U}$；（2）给定$\boldsymbol{U}$，优化参数$\boldsymbol{W}$。

（1）给定$\boldsymbol{W}$，更新$\boldsymbol{U}$: 包括如下子问题：

$$\min_{\boldsymbol{U}}\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle+\frac{\alpha_3}{2C}\|\boldsymbol{U}\|_F^2,\ \ s.t.\ \boldsymbol{U}\in\Gamma \tag{5.10}$$

注意到，定理5.2 仅给出了当$\alpha_3=0$时该问题的一个特例，无法对$\alpha_3\neq 0$的情况进行求解。下述定理表明，当α_3数值适中时，该子问题仍具有闭式解，且$\alpha_3=0$也符合该闭式解的形式，如图5.2所示。

定理 5.3. 令$\lambda_0(\mathcal{L}_{\mathcal{G}_{BI}})=0$，$\lambda_{N+1}(\mathcal{L}_{\mathcal{G}_{BI}})=+\infty$，记特征根$\lambda_1(\mathcal{L}_{\mathcal{G}_{BI}}),\cdots,\lambda_N(\mathcal{L}_{\mathcal{G}_{BI}})$对应的特征向量为$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_N]$。同时，记：

$$
\begin{aligned}
p &= \max\{i:\lambda_i(\mathcal{L}_{\mathcal{G}_{BI}})<\lambda_{i+1}(\mathcal{L}_{\mathcal{G}_{BI}}),0\le i<k\},\\
q &= \min\{i:\lambda_i(\mathcal{L}_{\mathcal{G}_{BI}})<\lambda_{i+1}(\mathcal{L}_{\mathcal{G}_{BI}}),i\ge k\},\\
\Delta p &= \lambda_{p+1}(\mathcal{L}_{\mathcal{G}_{BI}})-\lambda_p(\mathcal{L}_{\mathcal{G}_{BI}}),\\
\Delta q &= \lambda_{q+1}(\mathcal{L}_{\mathcal{G}_{BI}})-\lambda_q(\mathcal{L}_{\mathcal{G}_{BI}}),\\
\check{\delta}(\mathcal{L}_{\mathcal{G}_{BI}}) &= \min\{\Delta p,\Delta q\}
\end{aligned}
$$

对于任意$\mathcal{L}_{\mathcal{G}_{BI}}\neq\mathbf{0}$且满足$0\le\frac{\alpha_3}{2C}<\check{\delta}(\mathcal{L}_{\mathcal{G}_{BI}})$，问题(5.10)的最优解为：

$$
\boldsymbol{U}^\star=\boldsymbol{V}\tilde{\Lambda}\boldsymbol{V}^\top,\quad \tilde{\Lambda}=diag(\boldsymbol{c}),\quad c_i=\begin{cases}1 & i\le p,\\ \frac{k-p}{q-p} & q\ge i>p,\\ 0 & \text{otherwise.}\end{cases}\tag{5.11}
$$

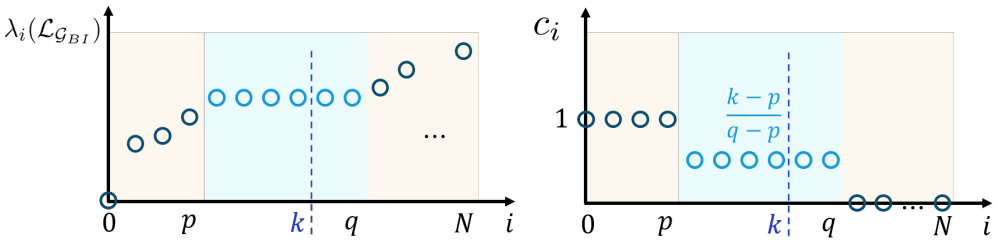


图 5.2 定理5.3示意图

Figure 5.2 Illustration of the Solution in Them. 5.3

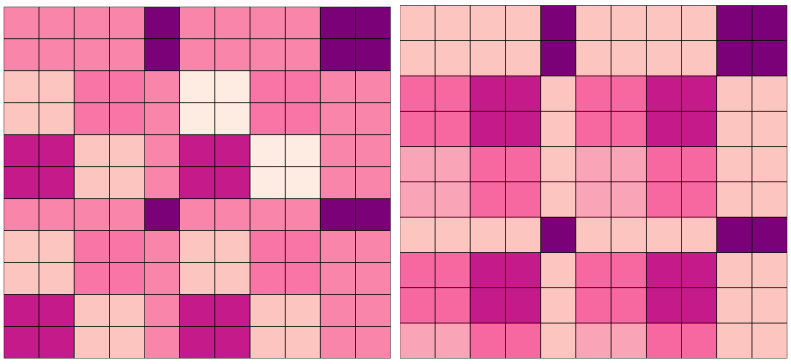
对于上述定理，有以下三个评述：

注 5.1 (理想条件下的分组效应). 首先讨论$\boldsymbol{V}$的分组能力。理想情况下，假设二部图有k个连通分量，由于$c_i=0,\forall i>k$，故仅$\boldsymbol{V}_{1:k}$与$\boldsymbol{U}^\star$的求解相关。记$\boldsymbol{V}_{1:k}$的第i行的转置为$\boldsymbol{f}_i\in\mathbb{R}^k$。接下来仅需考察$\boldsymbol{f}_i$的分组能力。定义$\mathcal{G}(1),\cdots,\mathcal{G}(k)$对应图中各个连通分量，其顶点数量分别为$n_{\mathcal{G}(1)},\cdots,n_{\mathcal{G}(k)}$。根据定理5.1，经过至多一次正交变换，$\boldsymbol{f}_i\in\mathbb{R}^{k\times1}$可表为：

$$
f_{i,j}=\begin{cases}\dfrac{1}{\sqrt{n_{\mathcal{G}(j)}}}, & i\in\mathcal{G}(j)\\ 0, & otherwise\end{cases}.\tag{5.12}
$$

由此看出，$\boldsymbol{f}_i$ 具有将任务/特征进行分组的强判别能力。第5.4.2.2小节通过详细的理论分析和更多的实际考量重新讨论此属性。

注 5.2. 与定理5.2仅考虑$\alpha_3 = 0$不同，定理5.3允许$\alpha_3 > 0$ 并赋予了$\boldsymbol{U}$子问题强凸性，使得算法具有全局收敛性。且能够证明算法全局收敛到原问题$(\boldsymbol{P})$与替代问题 $(\boldsymbol{P}^\star)$的临界点。下一小节将对此进行详细阐述。



(a) $\boldsymbol{V}_{1:2}\boldsymbol{V}_{1:2}^\top$ **(b)** $\boldsymbol{V}_{2:3}\boldsymbol{V}_{2:3}^\top$

图 5.3 向量外积示意图

Figure 5.3 Visualizations of the eigenvector outer-products

注 5.3. 记$\lambda_a(\mathcal{L}_{\mathcal{G}_{BI}}), \cdots, \lambda_b(\mathcal{L}_{\mathcal{G}_{BI}})$对应的特征向量为$\boldsymbol{V}_{a:b}$，如图5.2所示，即使特征根间隙$\lambda_{k+1}(\mathcal{L}_{\mathcal{G}_{BI}}) - \lambda_k(\mathcal{L}_{\mathcal{G}_{BI}})$接近于零，所提出算法同样有效。值得注意的是，若$\lambda_k(\mathcal{L}_{\mathcal{G}_{BI}}) = \lambda_{k+1}(\mathcal{L}_{\mathcal{G}_{BI}})$成立，则定理5.2的解没有明确定义。在该情形下，$\lambda_k(\mathcal{L}_{\mathcal{G}_{BI}})$ 的重数必定大于1。假设$\mathcal{L}_{\mathcal{G}_{BI}}$的最小$k$个特征根 $[\lambda_1(\mathcal{L}_{\mathcal{G}_{BI}}), \cdots \lambda_k(\mathcal{L}_{\mathcal{G}_{BI}})]$中有$s$个不同的特征根$[\check{\lambda}_1, \cdots \check{\lambda}_s]$ $(1 \le s < k)$，则$\boldsymbol{V}_{1:k}$无法张成子空间$\oplus_{i=1}^{s}\mathbb{EIG}_{\mathcal{L}_{\mathcal{G}_{BI}}}(\check{\lambda}_s)$（其中仅包含此子空间$q$个基中的$k$个）。可见$\boldsymbol{V}_{1:k}\boldsymbol{V}_{1:k}^\top$不能唯一确定（即，随正交变化及基的选择变化），因此解不具有可辨识性。换言之，选择不同的特征向量集合，可能产生完全不同的结果。例如，通过以下邻接矩阵构建一个二部图：

$$A = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{W} \\ \boldsymbol{W}^\top & \boldsymbol{0} \end{bmatrix}, \text{ 且 } \boldsymbol{W} = \begin{bmatrix} 1 & 1 & & & \\ 1 & 1 & & & \\ & & 2 & 2 & \\ & & 2 & 2 & \\ & & & & 3 \quad 3 \end{bmatrix}.$$

显然，该二部图的拉普拉斯矩阵的0特征根的重数为3，记对应的特征向量为$\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$。令$k = 2$，以下说明外积$\boldsymbol{V}\boldsymbol{V}^\top$不唯一。为此，首先对$\boldsymbol{V}_{1:2} = [\boldsymbol{v}_1, \boldsymbol{v}_2]$和 $\boldsymbol{V}_{2:3} =$

$[\boldsymbol{v}_2, \boldsymbol{v}_3]$分别计算其外积 $\boldsymbol{V}_{1:2}\boldsymbol{V}_{1:2}^{\top}$ 和 $\boldsymbol{V}_{2:3}\boldsymbol{V}_{2:3}^{\top}$。由图5.3的可视化结果可知，上述两个矩阵完全不同，进而产生完全不同的解。因此，上述子问题的定义并不明确。

与之相反，采用定理5.3则可规避该问题。基于定理5.3，有

$$\boldsymbol{U}^{\star} = \boldsymbol{V}_{1:p-1}\boldsymbol{V}_{1:p-1}^{\top} + \frac{k-p}{q-p}\boldsymbol{V}_{p:q}\boldsymbol{V}_{p:q}^{\top}.$$

进一步，根据q和p的定义可知，$\boldsymbol{V}_{p:q}$张成子空间$\mathbb{E}_1 = \mathbb{EIG}_{\mathcal{L}_{\mathcal{G}_{BI}}}(\breve{\lambda}_s)$，$\boldsymbol{V}_{1:p-1}$张成子空间$\mathbb{E}_2 = \bigoplus_{i=1}^{s-1}\mathbb{EIG}_{\mathcal{L}_{\mathcal{G}_{BI}}}(\breve{\lambda}_i)$。由此，$\boldsymbol{V}_{1:p-1}\boldsymbol{V}_{1:p-1}^{\top}$构成$\mathbb{E}_1$的正交投影，且$\boldsymbol{V}_{p:q}\boldsymbol{V}_{p:q}^{\top}$构成$\mathbb{E}_2$的正交投影。由正交投影的基本性质可知，$\boldsymbol{U}^{\star}$为旋转不变量。总结来说，相对于传统求解方法，定理5.3的优点如表5.1所示。注意此处的三种形式均产生相同的最优值，但有不同程度的稳定性。

表 5.1 $\sum_{i=1}^{k}\lambda_i(\mathcal{L}_{\mathcal{G}_{BI}})$**的不同变式**

Table 5.1 Different formulations of $\sum_{i=1}^{k}\lambda_i(\mathcal{L}_{\mathcal{G}_{BI}})$[1]

	凸	强凸	可辨识性
$\min_{\boldsymbol{V}} \ tr(\boldsymbol{V}\mathcal{L}_{\mathcal{G}_{BI}}\boldsymbol{V}^{\top})$ $s.t. \ \boldsymbol{V}\boldsymbol{V}^{\top} = \boldsymbol{I}_k$	×	×	×
$\min_{\boldsymbol{U}} \ \left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U} \right\rangle$ $s.t. \ \boldsymbol{U} \in \Gamma$	✔	×	×
$\min_{\boldsymbol{U}} \ \left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U} \right\rangle + \lambda \cdot \|\boldsymbol{U}\|_F^2$ $s.t. \ \boldsymbol{U} \in \Gamma$ $0 \le \lambda \le \breve{\delta}(\mathcal{L}_{\mathcal{G}_{BI}})$ $(Ours)$	✔	✔	✔

以下求解另一个子问题。

（2）固定$\boldsymbol{U}$，更新$\boldsymbol{W}$：下述命题表明，固定$\boldsymbol{U}$，可将$\boldsymbol{W}$子问题转换为弹性网正则近端映射（elastic net proximal mapping）问题：

命题 5.1. $(\boldsymbol{Prox})$的$\boldsymbol{W}$子问题的最优解为：

$$\boldsymbol{W}^{\star} = sgn(\widetilde{\boldsymbol{W}})\left(\left|\frac{\widetilde{\boldsymbol{W}}}{1+\frac{\alpha_2}{C}}\right| - \frac{\alpha_1}{C+\alpha_2}\boldsymbol{D}\right)_{+}, \tag{5.13}$$

[1]可辨识性系指当$\lambda_{k+1}(\mathcal{L}_{\mathcal{G}_{BI}}) - \lambda_k(\mathcal{L}_{\mathcal{G}_{BI}})$值为零时，$\boldsymbol{U} = \boldsymbol{V}\boldsymbol{V}^{\top}$的可辨识性。

其中， $D_{ij} = \|\boldsymbol{f}_i - \boldsymbol{f}_{d+j}\|^2$。

证明. 基于事实

$$\begin{aligned}
&\left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U} \right\rangle \\
&= \left\langle diag(\begin{bmatrix} 0 & |\boldsymbol{W}| \\ |\boldsymbol{W}|^\top & 0 \end{bmatrix}\mathbf{1}) - \begin{bmatrix} 0 & |\boldsymbol{W}| \\ |\boldsymbol{W}|^\top & 0 \end{bmatrix}, \boldsymbol{U} \right\rangle \\
&= \left\langle diag(\boldsymbol{U})\mathbf{1}^\top - \boldsymbol{U}, \begin{bmatrix} 0 & |\boldsymbol{W}| \\ |\boldsymbol{W}|^\top & 0 \end{bmatrix} \right\rangle,
\end{aligned} \tag{5.14}$$

该问题可重写为:

$$\min_{\boldsymbol{W}} \left\{ \begin{array}{l} \frac{1}{2}||\boldsymbol{W} - \widetilde{\boldsymbol{W}}||_F^2 + \frac{\alpha_1}{C} \cdot \left\langle \Delta^{(1)} + \Delta^{(2)^\top}, |\boldsymbol{W}| \right\rangle \\ + \frac{\alpha_2}{2C} \cdot \|\boldsymbol{W}\|_F^2 \end{array} \right\}, \tag{5.15}$$

其中，

$$\Delta = diag(\boldsymbol{U})\mathbf{1}^\top - \boldsymbol{U}, \tag{5.16}$$

$$\Delta^{(1)} = \Delta(1:d, (d+1):end), \tag{5.17}$$

$$\Delta^{(2)} = \Delta((d+1):end, 1:d). \tag{5.18}$$

进一步，有

$$\Delta_{ij}^{(1)} + \Delta_{ji}^{(2)} = U_{ii} + U_{d+j,d+j} - U_{i,d+j} - U_{d+j,i} = \|\boldsymbol{f}_i - \boldsymbol{f}_{d+j}\|_2^2.$$

鉴于目标函数是$\left(1 + \frac{\alpha_2}{C}\right)$−强凸的，易知，最优解唯一。故其证明遵循$\ell_1$范数的近端映射(Beck 等, 2009)。 □

给定嵌入向量，算法通过稀疏性诱导策略学习$\boldsymbol{W}$，其中，当$\widetilde{W}_{ij}$大于特征i和任务j间距离时，W_{ij}非零。进一步，下述评述揭示了跨特征、跨任务间的迁移机制。

注 5.4. 公式(5.10)具有如下等价形式:

$$\underset{\boldsymbol{W}}{\operatorname{argmin}} \langle \boldsymbol{D}, |\boldsymbol{W}| \rangle \ \ s.t. \boldsymbol{W} \in \mathcal{B}_{c(\alpha)}(\boldsymbol{W}, \widetilde{\boldsymbol{W}}^t), \tag{5.19}$$

其中，$\mathcal{B}_{c(\alpha)} = \left\{ \boldsymbol{W} : \|\boldsymbol{W} - \widetilde{\boldsymbol{W}}^t\|_F^2 \le c(\alpha_1), \|\boldsymbol{W}\|_F^2 \le c(\alpha_2) \right\}$。由该重构可知， $\boldsymbol{W}$子问题可等价表示为特征集合及任务集合之间的正则化最优运输问题。其中，运

输代价为任务、特征之间的谱嵌入距离；正则化项对应$\boldsymbol{W}$的F范数球限制。同时考虑到谱嵌入的分组性质，当特征i与任务j属于同一组时，运输代价低，$\boldsymbol{W}_{i,j}$优先被激活；反之当特征i与任务j属于不同组时，运输代价大，$\boldsymbol{W}_{i,j}$则不易被激活。因此迁移优先在相似的任务-特征间进行，可有效规避负迁移。

5.4.2　理论分析

5.4.2.1　收敛性分析

由于本章优化问题为非凸问题，因此收敛性分析主要考察更新序列是否能够有收敛到局部最优解，其中局部最优解由临界点，即次梯度为0的点，刻画。

为取到问题($\boldsymbol{Prox}$)的临界点，在改变参考点$\widetilde{\boldsymbol{W}}^t$之前，需迭代优化$\boldsymbol{U}$ 和$\boldsymbol{W}$ 直到收敛。直观而言，需由双层循环实现优化: 外循环 负责改变参考点；给定参考点，内循环负责求解$\boldsymbol{W}^k$和$\boldsymbol{U}^k$。显然，内循环大大增加了计算成本，使得双层循环在实际意义上并不可行。然而，实际训练中发现，一轮内循环足以收敛到良好的解。因此考虑如算法6的单次内循环更新过程。接下来对算法 6的全局收敛性加以证明。

定理 5.4 (算法6对替代问题($\boldsymbol{P}^\star$)的**全局收敛性**). 令$\{\boldsymbol{W}^t, \boldsymbol{U}^t\}$表示算法6产生的解序列。假设$\mathcal{J}(\cdot)$ 为可定义函数(definable function, 定义见本章附录) 且$\mathcal{J}(\boldsymbol{W}) \geq \boldsymbol{0}$，同时其导数为$\rho$-Lipschitz 连续函数。当$C > \rho$， $0 < \alpha_3 < 2C \min_t \check{\delta}(\mathcal{L}^t_{\mathcal{G}_{BI}})$时，对任意有限可行初始解，以下结论成立：

(1) 参数迭代序列$\{\boldsymbol{W}^t, \boldsymbol{U}^t\}_t$ 收敛到问题($\boldsymbol{P}^\star$)的一个临界点$(\boldsymbol{W}^\star, \boldsymbol{U}^\star)$。

(2) 损失序列收敛到问题 ($\boldsymbol{P}^\star$)的损失的一个临界点 $(\boldsymbol{W}^\star, \boldsymbol{U}^\star)$。

(3) 对于任意 $t \in \mathbb{N}$，存在一个次梯度$\boldsymbol{g}_t$，使得当$T \to +\infty$时，$\frac{1}{T}(\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2)$ 以$\mathcal{O}(\frac{1}{T})$的收敛率收敛到0。

定理 5.5. (**算法 6关于原问题 $\boldsymbol{P}$的全局收敛性**)。 在与定理5.4的相同条件下，算法 6所产生的迭代序列$\{\boldsymbol{W}^t, \boldsymbol{U}^t\}_t$关于原问题 ($\boldsymbol{P}$)同样满足上述结论(1)-(3)。

注 5.5. 由定理5.3可知，定理5.5所定义的原问题具有全局收敛性质。然而，直接采用定理5.2中的经典结论时，该结论并不总成立。原因主要有两个：一方面，即使$\boldsymbol{W}^t$收敛到临界点，也难以保证$\boldsymbol{U}^t$ 序列的可辨识性（见注5.3）；另一方面，定理5.2不具备定理 5.3所描述的性质，难以满足全局收敛性所必需的充分下降条件（见附录）。相反，通过所提出的定理5.3，即使在求解替代问题的前提下，

算法 6 求解TFCL原问题($\boldsymbol{P}$)

输入: 数据集$\mathcal{S}$、α_1、 α_2、k和$C(C > \varrho)$。

输出: 解 $\boldsymbol{W}$、$\boldsymbol{U}$。

初始化$\boldsymbol{W}^0$、$\boldsymbol{U}^0 \in \Gamma$、$t = 1$。

repeat

U子问题:

给定$\boldsymbol{W}^{t-1}$，计算$\mathcal{L}_{\mathcal{G}_{BI}}$。

$\boldsymbol{V}^t \leftarrow \mathcal{L}_{\mathcal{G}_{BI}}$的特征向量。

$\boldsymbol{U}^t \leftarrow \boldsymbol{V}^t \widetilde{\Lambda} \boldsymbol{V}^{t\top}$(根据定理5.3)。

W子问题:

给定 $\boldsymbol{U}^t$，计算$\boldsymbol{R}^t$。

$\widetilde{\boldsymbol{W}}^t \leftarrow \boldsymbol{W}^{t-1} - (1/C) \cdot \nabla_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W}^{t-1})$。

$\boldsymbol{W}^t \leftarrow sgn(\widetilde{\boldsymbol{W}}^t) \left(\left| \frac{\widetilde{\boldsymbol{W}}^t}{1 + \frac{\alpha_2}{C}} \right| - \frac{\alpha_1}{C+\alpha_2} \boldsymbol{D}^t \right)_+$。

$t \leftarrow t + 1$。

until 收敛

$\boldsymbol{W} = \boldsymbol{W}^{t-1}$，$\boldsymbol{U} = \boldsymbol{U}^{t-1}$。

也同样可以保证原问题全局收敛至其临界点。进一步，只要α 满足

$$\alpha \in (0,\ 2C \min_t \breve{\delta}(\mathcal{L}^t_{\mathcal{G}_{BI}})]$$

则算法结果不会受到影响，因此无需对α_3进行额外调参。

5.4.2.2 任务特征分组效应

本小节说明所提出的算法如何区分模型权重中的任务和特征组。具体而言，有如下定理。

理想情况下，当$\sum_{i=1}^{k} \lambda_i(\mathcal{L}_{\mathcal{G}_{BI}}) = 0$时，由定理5.1可知$\mathcal{G}_{BI}$必为$k$-连通的，且

$$\boldsymbol{W}_{i,j} \neq 0$$

当且仅当特征 i和任务j属于同一连通分量。

实际上，$\sum_{i=1}^{k} \lambda_i(\mathcal{L}_{\mathcal{G}_{BI}})$往往大于零，因此上述论证并不成立。对于一个性能良好的模型，可以假设目标函数值在算法后期阶段较小。受此启发，以下定理表明，在上述更弱的假设条件下，仍可恢复成组结构。

定理 5.6. (**成组效应**) 假设算法 6 在第$\mathcal{T}$轮迭代停止，且$\mathcal{F}(\boldsymbol{W}^{\mathcal{T}-1},\boldsymbol{U}^{\mathcal{T}-1}) \leq \epsilon_{\mathcal{T}-1}$。记$\mathsf{Supp}(\boldsymbol{A}) = \{(i,j) : A_{i,j} \neq 0\}$，$\mathcal{H}_K = \{\boldsymbol{W} : \|\boldsymbol{W}\|_F \leq K\}$，$C_0 = \left(\frac{2}{\alpha_2} \cdot \epsilon_{\mathcal{T}-1}\right)^{1/2}$。另假设 $\infty > \kappa > 0$，$\sup_{\|\boldsymbol{W}\|_F \leq \kappa} \|\nabla_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W})\|_\infty \leq \varpi(\kappa) < \infty$ 且，有矩阵$\boldsymbol{W}^\star \in \mathcal{H}_{C_0}$，其对应的二部图$\mathcal{G}^\star$ 有k个连通分量，图拉普拉斯矩阵 $\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}}$给出了真实分组。此外，记图中第$i$个组的顶点数为$n_i$，记$n_1^{\uparrow} = \max_i n_i$，$n_2^{\uparrow} = \max_{j, n_j \leq n_1^{\uparrow}} n_i$。定义符号

$$\begin{aligned} &\kappa_0 = C_0 + \frac{\varpi(C_0)}{C},\ \delta_1 = \frac{C}{\alpha_1}\kappa_0,\ \delta_2 = \frac{C}{\alpha_1}\delta_0,\ \beta = \frac{1}{n_1^{\uparrow}} + \frac{1}{n_2^{\uparrow}}, \\ &\rho = \frac{C_0}{\lambda_{k+1}\left(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}}\right)},\ \xi = \rho \cdot (\sqrt{d+T} + \sqrt{2}), \end{aligned} \tag{5.20}$$

有：

(a) (**无假阳性分组**) 当 $\lambda_{k+1}(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}}) > \lambda_k(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}}) > 0$，$\frac{\sqrt{2}}{32} \cdot \beta > \xi$，且$8\sqrt{2}\xi < \delta_1 < \beta - 8\sqrt{2}\xi$时，有：

$$\mathsf{Supp}(\boldsymbol{W}^{\mathcal{T}}) \subseteq \{(i,j) : \mathcal{G}(i) = \mathcal{G}(j)\} = \mathsf{Supp}(\boldsymbol{W}^\star), \tag{5.21}$$

其中，$\mathcal{G}(i)$为二部图 $\mathcal{G}^\star$中顶点i对应的连通分量。

(b) (**正确分组**) 若进一步假设$\min_{(i,j)} |\widetilde{\boldsymbol{W}}^{\mathcal{T}}_{i,j}| \geq \delta_0 > 0$，且$8\sqrt{2}\xi < \min\{\delta_1, \delta_2\} \leq \max\{\delta_1, \delta_2\} < \beta - 8\sqrt{2}\xi$，有：

$$\mathsf{Supp}(\boldsymbol{W}^{\mathcal{T}}) = \mathsf{Supp}(\boldsymbol{W}^\star). \tag{5.22}$$

注 5.6. 对于定理，作出如下解释：

(a) 假设总存在$\boldsymbol{W}^\star \in \mathcal{H}_{C_0}$，满足与真实成组参数结构一致（反之，若$\boldsymbol{W}^\star \notin \mathcal{H}_{C_0}$，可选择$\boldsymbol{W}' = C_0 \cdot \frac{\boldsymbol{W}^\star}{||\boldsymbol{W}^\star||_F}$，而$\boldsymbol{W}'$ 位于同一支撑集的F-范数球内）。

(b) 定理5.6说明，若$\mathcal{L}_{\mathcal{G}_{BI}}$的第$k$个谱间隙存在，选取超参数：

$$\alpha_2 = o\left(\frac{(d+T) \cdot \epsilon_{\mathcal{T}-1}}{\beta^2 \cdot \lambda_{k+1}(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}})^2}\right),\ \alpha_1 = \mathcal{O}(C\kappa_0) \tag{5.23}$$

并确保关于ξ, δ_1的不等式成立，则可实现无假阳性结构恢复，其中，仅当特征 i 和任务 j在真实参数结构中属于同一组时，$|W^{\mathcal{T}}_{ij}|$值为零。进一步假设：

$$\alpha_1 = \mathcal{O}(C \cdot (\delta_0 \vee \kappa_0)),\ \alpha_2 = o\left(\frac{(d+T) \cdot \epsilon_{\mathcal{T}-1}}{\beta^2 \cdot \lambda_{k+1}(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}})^2}\right) \tag{5.24}$$

其中，$a \vee b$ 表示取二者中较大值，则可近似获得正确的分组结构。换言之，仅当特征 i 和任务 j在真实参数结构中属于同一组时，$|W^{\mathcal{T}}_{ij}|$值为零。

5.4.3 与既往工作关系

本小节讨论所提出的基本模型与既往工作(Lu 等, 2019)和(Xu 等, 2015)间的关系。与(Xu 等, 2015)类似，所提出的模型也假设应将特征和任务划分到不同的组，而不同之处在于：工作(Xu 等, 2015)基于k-均值假设进行协同聚类，无法保证块对角结构。受(Lu 等, 2019)启发，本章借鉴谱图论为聚类设计了一个显式的正则化算子，将(Lu 等, 2019)中的正则化项算子推广到二部图上并将其应用于多任务学习问题。更重要的是，还为定理5.3中的截断特征根之和问题的变式提供了一个广义闭式解。值得注意的是，该问题恢复了原问题（非强凸）的一个形如式 (5.10)的特定全局解，因此同时兼顾了唯一性与最优性。

5.4.4 个性化属性预测

至此，已介绍了所提出的任务-特征协同学习框架，以下将此框架拓展到个性化属性预测问题上。给定大量来自于众包标注平台的图像及其视觉属性标注（例如，人脸的微笑属性，鞋子的舒适性），个性化属性预测旨在给未标注图像标注用户特异的视觉属性，以满足复杂的个性化需求。由于通常每个众包用户仅标注少量图像，该问题与多任务学习问题具有天然的相似性。

为建模个性化标注过程，视每个用户的标注预测为一个任务。对于一个给定的属性，假设有T个用户参与了该属性的标注。同时，假设第i个用户标注了n_i张图片，其中，有$n_{+,i}$个正标签，$n_{-,i}$个负标签。在此设定下，$\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_i \times d}$ 为第i个用户所标注图片的输入特征，$\boldsymbol{y}^{(i)} \in \{-1, 1\}^{n_i}$为对应的标签向量。$y_k^{(i)} = 1$表示用户认为第$s$张图片中出现了目标属性，否则，$y_k^{(i)} = -1$。此外，令$\mathcal{S}_{+,i} = \{k \mid y_k^{(i)} = 1\}$，$\mathcal{S}_{-,i} = \{k \mid y_k^{(i)} = -1\}$。多样的个性化标注允许为每个用户进行不同建模。受此启发，为每个任务(用户) i建立了一个线性学习器$\boldsymbol{g}^{(i)}(\boldsymbol{x}) = \boldsymbol{W}^{(i)\top}\boldsymbol{x}$。与第4章相似，本章将参数$\boldsymbol{W}$分解为：

$$\boldsymbol{W}^{(i)} = \boldsymbol{\Theta}_c + \boldsymbol{\Theta}_g^{(i)} + \boldsymbol{\Theta}_p^{(i)}. \tag{5.25}$$

其中$\boldsymbol{\Theta}_c$、$\boldsymbol{\Theta}_g^{(i)}$和$\boldsymbol{\Theta}_p^{(i)}$的含义与第4章中$\boldsymbol{\theta}$、$\boldsymbol{G}^{(i)}$和$\boldsymbol{P}^{(i)}$相同，分别捕捉全局、群体与个性化属性偏好。同理，记$\boldsymbol{\Theta}_g = [\boldsymbol{\Theta}_g^{(1)}, \cdots, \boldsymbol{\Theta}_g^{(T)}]$，$\boldsymbol{\Theta}_p = [\boldsymbol{\Theta}_p^{(1)}, \cdots, \boldsymbol{\Theta}_p^{(T)}]$。

为实现有效的参数分解，需为各参数设计对应正则化算子。与第4章相同，针对$\boldsymbol{\Theta}_c$使用ℓ_2正则化算子；针对$\boldsymbol{\Theta}_g$使用所提出的任务-特征协同学习框架；针对$\boldsymbol{\Theta}_p$采用$\ell_{1,2}$正则化算子；使用AUC作为优化目标，且选择平方替代损失(Gao

等, 2016)：

$$\begin{aligned}\mathcal{J}(\boldsymbol{\Theta}_c,\boldsymbol{\Theta}_g,\boldsymbol{\Theta}_p) &= \sum_{i=1}^{T}\ell_i,\\ \ell_i &= \sum_{x_p\in\mathcal{S}_{+,i}}\sum_{x_q\in\mathcal{S}_{-,i}}\frac{s\left(\boldsymbol{g}^{(i)}(\boldsymbol{x}_p)-\boldsymbol{g}^{(i)}(\boldsymbol{x}_q)\right)}{n_{+,i}n_{-,i}}.\end{aligned} \tag{5.26}$$

其中$s(t) = (1-t)^2$。

综上所述，目标函数有以下形式：

$$(\boldsymbol{Q})\ \min_{\boldsymbol{\Theta},\boldsymbol{U}\in\Gamma}\left\{\begin{array}{l}\mathcal{J}(\boldsymbol{\Theta}_c,\boldsymbol{\Theta}_g,\boldsymbol{\Theta}_p)+\frac{\alpha_1}{2}\|\boldsymbol{\Theta}_c\|_2^2+\ \alpha_2\ \langle\boldsymbol{\Theta}_{\mathcal{G}},\ \boldsymbol{U}\rangle\\ +\frac{\alpha_3}{2}\ \|\boldsymbol{\Theta}_g\|_F^2\ +\ \alpha_4\ \|\boldsymbol{\Theta}_p\|_{1,2}\ +\ \iota_\Gamma(U)\end{array}\right\}. \tag{5.27}$$

优化 可证明，当输入$\boldsymbol{X}$有界时，经验损失$\mathcal{J}=\sum_{i=1}^{T}\ell_i$关于$\boldsymbol{\Theta}=[\boldsymbol{\Theta}_c, vec(\boldsymbol{\Theta}_g), vec(\boldsymbol{\Theta}_p)]$的导数是Lipschitz连续的。记Lipschitz常数为$\varrho_{\boldsymbol{\Theta}}$ (见附录)。任意第t此迭代取$C > \varrho_{\boldsymbol{\Theta}}$，可通过如下子问题求解参数：

$$(\boldsymbol{Prox_g})\ \operatorname*{argmin}_{\boldsymbol{\Theta}_g,\boldsymbol{U}\in\Gamma}\left\{\begin{array}{l}\frac{1}{2}\left\|\boldsymbol{\Theta}_g-\widetilde{\boldsymbol{\Theta}}_g^t\right\|_F^2+\frac{\alpha_1}{C}\langle\boldsymbol{\Theta}_{\mathcal{G}},\boldsymbol{U}\rangle\\ +\frac{\alpha_2}{2C}\|\boldsymbol{\Theta}_g\|_F^2\end{array}\right\}, \tag{5.28}$$

$$(\boldsymbol{Prox_c})\ \operatorname*{argmin}_{\boldsymbol{\Theta}_c}\frac{1}{2}\left\|\boldsymbol{\Theta}_c-\widetilde{\boldsymbol{\Theta}}_c^t\right\|_2^2+\frac{\alpha_3}{2C}\|\boldsymbol{\Theta}_c\|_2^2, \tag{5.29}$$

$$(\boldsymbol{Prox_p})\ \operatorname*{argmin}_{\boldsymbol{\Theta}_p}\frac{1}{2}\left\|\boldsymbol{\Theta}_p-\widetilde{\boldsymbol{\Theta}}_p^t\right\|_F^2+\frac{\alpha_4}{C}\|\boldsymbol{\Theta}_p\|_{1,2}, \tag{5.30}$$

其中，

$$\widetilde{\boldsymbol{\Theta}}_g^t=\boldsymbol{\Theta}_g^{t-1}-\frac{1}{C}\nabla_{\boldsymbol{\Theta}_g}\mathcal{J}(\boldsymbol{\Theta}^{t-1}), \tag{5.31}$$

$$\widetilde{\boldsymbol{\Theta}}_c^t=\boldsymbol{\Theta}_c^{t-1}-\frac{1}{C}\nabla_{\boldsymbol{\Theta}_c}\mathcal{J}(\boldsymbol{\Theta}^{t-1}), \tag{5.32}$$

$$\widetilde{\boldsymbol{\Theta}}_g^t=\boldsymbol{\Theta}_p^{t-1}-\frac{1}{C}\nabla_{\boldsymbol{\Theta}_p}\mathcal{J}(\boldsymbol{\Theta}^{t-1}). \tag{5.33}$$

与算法6相似，采用算法7优化参数。

本节的最后说明，算法7继承了算法6的理论优势。

定理 5.7. 记损失函数为

$$\widetilde{\mathcal{F}} = \left\{ \begin{array}{l} \mathcal{J}(\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_g, \boldsymbol{\Theta}_p) + \alpha_1 \langle \boldsymbol{\Theta}_{\mathcal{G}}, \boldsymbol{U} \rangle + \dfrac{\alpha_2}{2} \|\boldsymbol{\Theta}_g\|_F^2 \\ + \dfrac{\alpha_3}{2} \|\boldsymbol{\Theta}_c\|_2^2 + \alpha_4 \|\boldsymbol{\Theta}_p\|_{1,2} + \iota_{\Gamma}(\boldsymbol{U}) \end{array} \right\}$$

记第t轮迭代产生的参数为$(\boldsymbol{\Theta}_c^t, \boldsymbol{\Theta}_g^t, \boldsymbol{\Theta}_p^t, \boldsymbol{U}^t)$。若输入特征有界，即对于$\forall i$, $\|\boldsymbol{X}^{(i)}\|_2 = \vartheta_{X_i} < \infty$, $n_{+,i} \geq 1$, $n_{-,i} \geq 1$， $C > \varrho_{\Theta}$，其中 $\varrho_{\boldsymbol{\Theta}} = 3T\sqrt{(2T+1)} \max_i \left\{ \dfrac{n_i \vartheta_{X_i}^2}{n_{+,i} n_{-,i}} \right\}$，下列性质对算法7而言成立：

(1) 参数迭代序列$(\boldsymbol{\Theta}_c^t, \boldsymbol{\Theta}_g^t, \boldsymbol{\Theta}_p^t, \boldsymbol{U}^t)$ 收敛到问题$\boldsymbol{Q}$的一个临界点 $(\boldsymbol{\Theta}_c{}^{\star}, \boldsymbol{\Theta}_g{}^{\star}, \boldsymbol{\Theta}_p{}^{\star}, \boldsymbol{U}^{\star})$。

(2) 损失序列$\{\widetilde{\mathcal{F}}_t\}_t$ 收敛到问题 $\boldsymbol{Q}$的一个临界点 $\widetilde{\mathcal{F}}^{\star}$。

(3) 对于任意$t \in \mathbb{N}$，存在一个次梯度 $\boldsymbol{g}_t$，从而当$T \to +\infty$时， $\dfrac{1}{T}(\sum\limits_{t=1}^{T} \|\boldsymbol{g}_t\|^2)$以$\mathcal{O}(\frac{1}{T})$的收敛率收敛到0。

定理 5.8. 假设算法 7 迭代到第$\mathcal{T}$轮终止，且满足$\tilde{\mathcal{F}}_{\mathcal{T}-1} \leq \epsilon_{\mathcal{T}-1}^{\mathcal{A}}$，其中$\tilde{\mathcal{F}}_{\mathcal{T}-1}$为第$\mathcal{T}-1$轮迭代时目标函数的值。假设存在矩阵$\boldsymbol{\Theta}_{\boldsymbol{g}}{}^{\star} \in \mathcal{H}_{C_0^{\mathcal{A}}}$，其对应的二部图 $\mathcal{G}^{\star}$有k个连通分量，且图拉普拉斯矩阵 $\boldsymbol{\Theta}_{\mathcal{G}}^{\mathcal{T}}$给出了真实分组结构。此外，记 n_i为二部图第i个连通分量对应的顶点个数，且 $n_1^{\uparrow} = \max_i n_i$， $n_2^{\uparrow} = \max_{j, n_j \leq n_1^{\uparrow}} n_i$。基于下列符号定义：

$$C_0^{\mathcal{A}} = \left(2 \cdot \frac{\epsilon_{\mathcal{T}-1}^{\mathcal{A}}}{\alpha_2}\right)^{1/2}, \kappa_0^{\mathcal{A}} = C_0^{\mathcal{A}} + \frac{\varkappa(\xi_c, \xi_g, \xi_p)}{C}, \delta_1^{\mathcal{A}} = \frac{C}{\alpha_1} \kappa_0^{\mathcal{A}} \tag{5.34}$$

$$\delta_2^{\mathcal{A}} = \frac{C}{\alpha_1} \delta_0^{\mathcal{A}}, \ \ \xi^{\mathcal{A}} = (\sqrt{d+T} + \sqrt{2}) \cdot \frac{C_0^{\mathcal{A}}}{\lambda_{k+1}(\boldsymbol{\Theta}_{\mathcal{G}}^{\mathcal{T}})} \tag{5.35}$$

$$\begin{aligned} &\xi_c = \left(2 \cdot \frac{\epsilon_{\mathcal{T}-1}^{\mathcal{A}}}{\alpha_3}\right)^{1/2}, \ \xi_g = C_0^{\mathcal{A}}, \ \xi_p = \frac{\epsilon_{\mathcal{T}-1}^{\mathcal{A}}}{\alpha_4}, \ \beta^{\mathcal{A}} = \frac{1}{n_1^{\uparrow}} + \frac{1}{n_2^{\uparrow}}, \\ &\varkappa(\xi_c, \xi_g, \xi_p) = \frac{n_i \vartheta_{X_i}}{\sqrt{n_{+,i}} n_{-,i}} \sum_{i=1}^{T} \left((\xi_c + \xi_g + \xi_p) \frac{\vartheta_{X_i}}{\sqrt{n_{+,i}}} + 1 \right), \end{aligned} \tag{5.36}$$

对于算法7中$\boldsymbol{\Theta}_g$的分组效应，有下述事实成立:

(a) **无假阳性分组**：$\lambda_{k+1}(\boldsymbol{\Theta}_{\mathcal{G}}^{\mathcal{T}}) > \lambda_k(\boldsymbol{\Theta}_{\mathcal{G}}^{\mathcal{T}}) \geq 0$, $\frac{\sqrt{2}}{32} \cdot \beta^{\mathcal{A}} > \xi^{\mathcal{A}}$，且 $8\sqrt{2}\xi^{\mathcal{A}} < \delta_1^{\mathcal{A}} < \beta^{\mathcal{A}} - 8\sqrt{2}\xi^{\mathcal{A}}$有：

$$\mathrm{Supp}(\boldsymbol{\Theta}_g^{\mathcal{T}}) \subseteq \left\{(i,j) : \mathcal{G}(i) = \mathcal{G}(j)\right\} = \mathrm{Supp}(\boldsymbol{\Theta}_g^{\star}), \tag{5.37}$$

其中，$\mathcal{G}(i)$为相应二部图 $\mathcal{G}^{\star}$中节点i所属的连通分量。

(b) **正确分组**：若进一步假设 $\min_{(i,j)} |\widetilde{\boldsymbol{\Theta}}_{i,j}^{\mathcal{T}}| \geq \delta_0^{\mathcal{A}} > 0$，$8\sqrt{2}\xi^{\mathcal{A}} < \min\left\{\delta_1^{\mathcal{A}}, \delta_2^{\mathcal{A}}\right\} \leq \max\left\{\delta_1^{\mathcal{A}}, \delta_2^{\mathcal{A}}\right\} < \beta^{\mathcal{A}} - 8\sqrt{2}\xi^{\mathcal{A}}$，有:

$$\mathrm{Supp}(\boldsymbol{\Theta}_g^{\mathcal{T}}) = \mathrm{Supp}(\boldsymbol{\Theta}_g^{\star}). \tag{5.38}$$

算法 7 ($\boldsymbol{Q}$)问题求解

输入： 数据集 $\mathcal{S}$、α_1、α_2、α_3、α_4、k、$C(C > \varrho_{\boldsymbol{\Theta}})$。

输出： 解 $\boldsymbol{\Theta}_c$、$\boldsymbol{\Theta}_g$、$\boldsymbol{\Theta}_p$、$\boldsymbol{U}$。

初始化$\boldsymbol{\Theta}_c^0$、 $\boldsymbol{\Theta}_g^0$、$\boldsymbol{\Theta}_p^0$、 $\boldsymbol{U}^0 \in \Gamma$、$t = 1$。

repeat

根据公式(5.32)-公式(5.33)，分别计算$\widetilde{\boldsymbol{\Theta}}_c^t$，$\widetilde{\boldsymbol{\Theta}}_g^t$和 $\widetilde{\boldsymbol{\Theta}}_g^t$。

根据公式(5.29)求解$\boldsymbol{\Theta}_c^t$。

根据公式(5.30)求解$\boldsymbol{\Theta}_p^t$。

基于$\mathcal{S}$、$\alpha_1 = \alpha_2, \alpha_2 = \alpha_3$、$k$、$C$，调用算法6并返回$\boldsymbol{\Theta}_g^t, \boldsymbol{U}^t$。

$t = t + 1$。

until 收敛

$\boldsymbol{\Theta}_c = \boldsymbol{\Theta}_c^{t-1}$，$\boldsymbol{\Theta}_g = \boldsymbol{\Theta}_g^{t-1}$，$\boldsymbol{\Theta}_p = \boldsymbol{\Theta}_p^{t-1}$，$\boldsymbol{U} = \boldsymbol{U}^{t-1}$。

5.5　实验

5.5.1　数据集

5.5.1.1　仿真数据集

为了测试TFCL基本模型的有效性，采用100个仿真用户生成一个简单的仿真标注数据集，其中特征与AUC分数由带有一个块对角任务矩阵的线性模型生成。对于每一个用户，生成200个样品，得到输入特征矩阵$\boldsymbol{X}^{(i)} \in \mathbb{R}^{200\times 80}$，其中$\boldsymbol{x}_k^{(i)} \sim \mathbb{N}(0, \boldsymbol{I}_{80})$。本章按照下述方式生成块对角任务矩阵$\boldsymbol{W}$。具体而言，按照公式$\boldsymbol{W} = \bigoplus_{i=1}^{5} \boldsymbol{W}_i$ 构造5个对角块，其中 $\boldsymbol{W}_1 \in \mathbb{R}^{20\times 20}$，$\boldsymbol{W}_2 \in \mathbb{R}^{20\times 20}$，$\boldsymbol{W}_3 \in \mathbb{R}^{10\times 20}$，$\boldsymbol{W}_4 \in \mathbb{R}^{20\times 20}$，$\boldsymbol{W}_5 \in \mathbb{R}^{10\times 20}$为$\boldsymbol{W}$。每个对角块中的元素采样于分布$\mathbb{N}(C_i, 2.5^2)$（通

过逐元素采样的方式生成），其中，$C_i \sim \mathbb{U}(0, K_i), K_1 = 5, K_2 = 5, K_3 = 10, K_4 = 15, K_5 = 20$对应相应的聚类中心。对于每个用户，打分函数为$\boldsymbol{s}^{(i)} = \boldsymbol{X}^{(i)}(\boldsymbol{W}^{(i)} + \epsilon^{(i)})$，其中$\epsilon^{(i)} \in \mathbb{R}^{200\times 1}$且$\epsilon^{(i)} \sim \mathbb{N}(0, 0.1^2\boldsymbol{I}_{200})$。接下来为每一个任务$i$生成相应标签$\boldsymbol{Y}^{(i)}$，具体而言，将得分最高的50个任务实例标定为1，其余则标定为-1。

5.5.1.2 Shoes数据集

Shoes数据集(Kovashka 等, 2015)是流行的属性预测基准数据集，包含对应7种属性的14,658幅在线购物图片（BR：棕色，CM：舒适的，FA：时尚的，FM：正式的，OP：打开的，ON：华丽的，PT：尖的）。具有不同知识背景的用户需判断图片中是否存在特定属性。具体而言，每类属性至少分配了190名用户参与标注，而每位用户均标注了50幅图片，共计获得90,000份标注结果。

5.5.1.3 Sun数据集

Sun(Patterson 等, 2012)数据集包含来自Sun属性数据库 (Patterson 等, 2012)的14, 340幅场景图片。通过相似的标注过程，共计获得包含5种属性的64, 900份个性化标注结果（CL：杂乱无章的，MO：现代的，OP：开放领域，RU：乡村的，SO：平静的）。

5.5.2 对比方法

接下来简要介绍实验涉及的对比方法。

- **LASSO** (Tibshirani, 1996)，在该方法中，每一个任务学习器均采用ℓ_1-范数进行正则化约束。
- **rMTFL** (Yu 等, 2007)，该方法假定模型参数$\boldsymbol{W}$可被分解为共识组件和成组稀疏组件两部分。
- **RAMUSA** (Han 等, 2016)，该方法采用capped迹范数正则化算子，从而仅最小化小于自适应调整阈值的奇异值。
- **CoCMTL** (Xu 等, 2015)，该方法通过最小化任务矩阵的奇异值的截断平方和来实现任务特异的协同聚类。
- **NC-CMTL** (Nie 等, 2018)，该方法使用非凸的低秩谱正则化算子和重加权方案来探索不同任务间的共享信息。
- **VSTGMTL** (Jeong 等, 2018)，该方法同时进行变量选择和低秩分解学习。
- **AMTL** (Lee 等, 2016) 通过在非对称迁移矩阵上进行稀疏选择来实现跨任

务间的非对称迁移。

5.5.3　实验细节

计算所有用户上AUC分数的均值作为评价标准。对所有数据集独立进行了15 次训练集、验证集、测试集的采集，并根据模型在验证集（占总样本数的85%）上15次结果的均值调整超参数，记录在测试集上15次结果的均值。对于对比方法，基于已有的预测结果 $[\hat{\boldsymbol{y}}^{(1)}, \cdots, \hat{\boldsymbol{y}}^{(T)}]$，实验采用了逐样本平方损失之和作为最终的损失函数，即：

$$\sum_i \frac{1}{2} \cdot ||\boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}||_2^2$$

对于本章所提出的算法，采用平方AUC损失函数以取得更好性能。为公平起见，同时采用逐样本平方损失之和作为本章所提出算法的最终损失函数，并记录实验结果进行比较。

5.5.4　实验结果

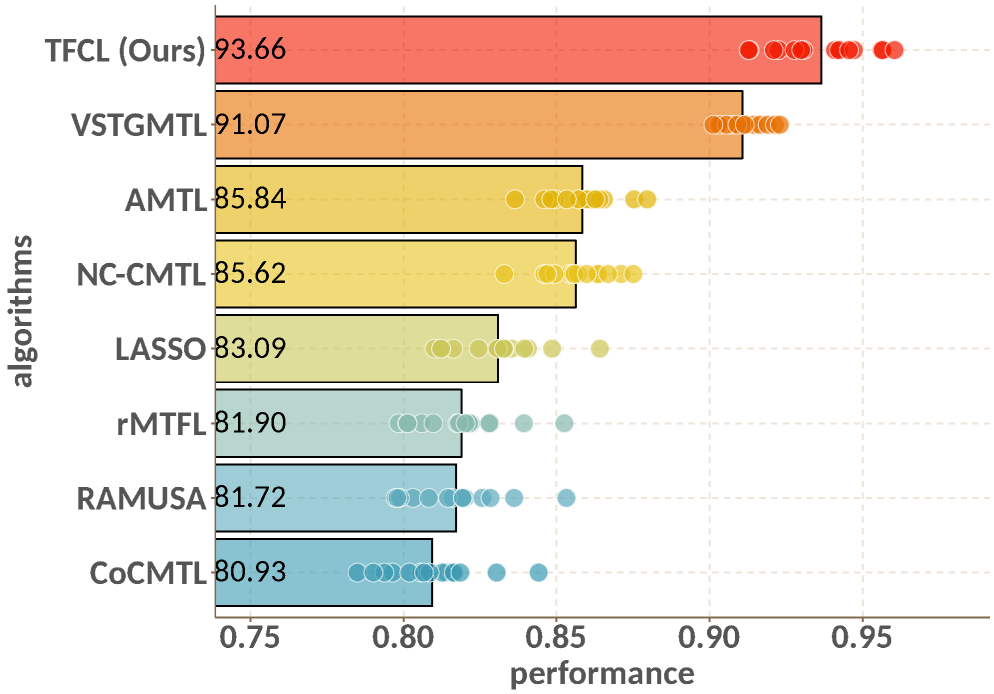
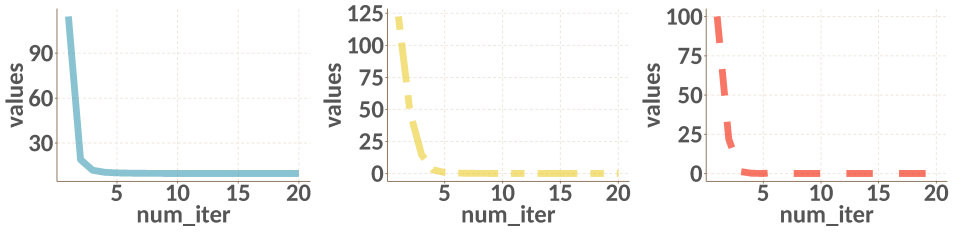


图 5.4 在仿真数据集上的AUC (↑)消融实验结果示意图。

Figure 5.4 AUC (↑) comparison on the Simulated Dataset

[2]即$||\boldsymbol{W}^t - \boldsymbol{W}^{t-1}||$

[3]y轴代表测试集上的平均AUC分数，x轴代表不同的算法，分别是：**Org**展示了初始TFCL算法的表现；



(a) obj (b) $||W^t - W^{t-1}||$ (c) $||U^t - U^{t-1}||$

图 5.5 （a）损失函数收敛曲线；（b）参数变化收敛曲线

Figure 5.5 (a) Convergence curves for loss function; (b) Convergence curves for parameter change[2]

表 5.2 仿真数据集上的消融研究

Table 5.2 Ablation Study for simulated dataset

Algorithm	TFCL ours	TFCL w/o AUC loss
AUC	**93.66**	92.46

5.5.4.1 仿真数据集

性能对比： 图5.4展示仿真数据集上所有相关算法的性能。相关结果表明，所提出的算法均优于对比方法。在15次重复实验中，所提出的方法AUC性能达到93.66%，相较于次优方法的91.07%提高了2.59%。

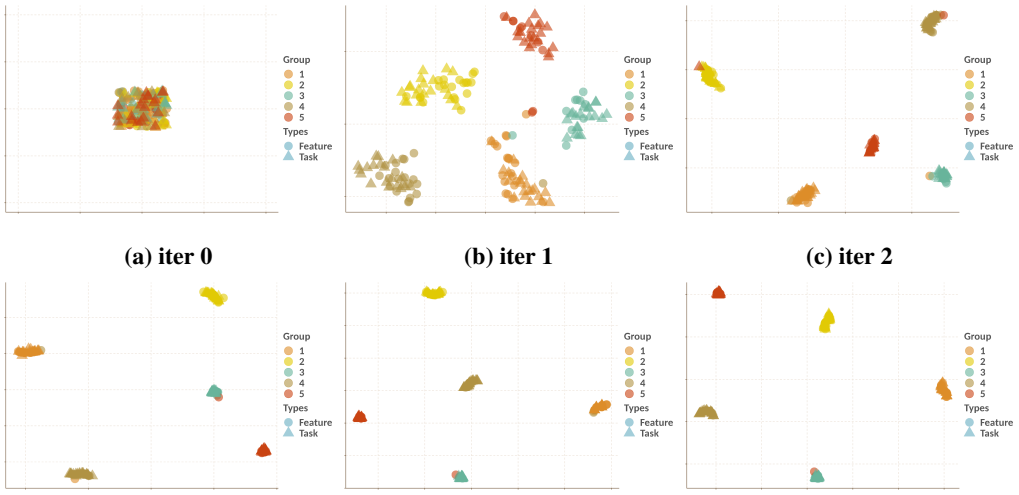
消融实验： 表5.2展示消融实验结果，以分别说明AUC损失和群体因素对模型性能的影响。通过将原模型中AUC损失替换为逐样本加权均方损失得到基线模型，可发现：(1) 原模型优于基线模型，即采用AUC优化是有效的；(2) 基线模型优于其他对比方法，即群体因素在仿真数据集上比其他对比方法有效。

收敛分析： 如图5.5a和5.5c所示，所提出的方法在目标函数和参数迭代序列上均具有良好的收敛性，和理论结果相吻合。

谱嵌入可视化： 图5.6包含前5轮迭代中谱嵌入的演化过程。结果显示，前

w/o_AUC展示了使用平方损失函数作为AUC替代损失函数时算法的表现；**w/o_G**展示了移除联合分组参数时算法的表现。

[4] y轴代表测试集上的平均AUC分数，x轴代表不同的算法：TFCL(Coc)表示TFCL算法在使用CocMTL中对应的正则化算子替代所提出的联合分组参数后的表现；TFCL(ours)表示初始的TFCL算法的表现。



(a) iter 0　(b) iter 1　(c) iter 2

(d) iter 3　(e) iter 4　(f) iter 5

图 5.6 谱嵌入的演化

Figure 5.6 Evolution of Spectral Embeddings

两轮迭代即可形成稳定清晰的聚类，验证了关于谱嵌入分组能力的理论分析。联系图5.5和图5.6，可发现参数迭代曲线和嵌入空间演化间存在紧密的联系，即在算法6中谱嵌入与$\boldsymbol{V}$在同一个子问题中优化，从而最小化损失函数。如图5.5所示，损失函数快速下降并于第5轮迭代收敛，而在图5.6中嵌入同样于5轮迭代内收敛至对应的簇。

结构恢复： 为验证所提出模型恢复参数$\boldsymbol{W}$预期结构的能力，图5.7比较了相同仿真数据集上相关算法参数$\boldsymbol{W}$与真实结构的差别。结果显示，所提出算法能够恢复更清晰的参数结构。同时，所有的对比方法均可以大致恢复块对角轮廓，但存在不同程度的非对角噪声。通过算术分析理解，线性模型的判别函数为$\hat{\boldsymbol{y}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}$，而真实参数通过线性函数$\boldsymbol{X}\boldsymbol{W}^{\star}$生成。若矩阵$\boldsymbol{X}$不满秩，则每当$\boldsymbol{W} = \boldsymbol{W}^{\star} + \boldsymbol{W}'$和$\boldsymbol{W}' \in \mathrm{Null}(\boldsymbol{X})$有$\boldsymbol{X}\boldsymbol{W} = \boldsymbol{X}\boldsymbol{W}^{\star}$。$\boldsymbol{W}'$通常存在非对角元素，因此自然产生如图5.7所示的噪声。若不使用块对角正则化算子，即使全局最优解也难以避免$\boldsymbol{W}'$。此外，由于$\boldsymbol{W}'$仅和可观测数据$\boldsymbol{X}$相关，因此存在过拟合风险，尤其是在非对角噪声与仿真数据集存在冲突的情况。通过消除非对角噪声，所提出的方法取得了显著的性能提升。

5.5.4.2　Shoes数据集

预处理： 将原始数据集提供的GIST(Kovashka 等, 2012)和颜色直方图串联以生成输入特征。接着，在执行训练前通过主成分分析法（Principal Component

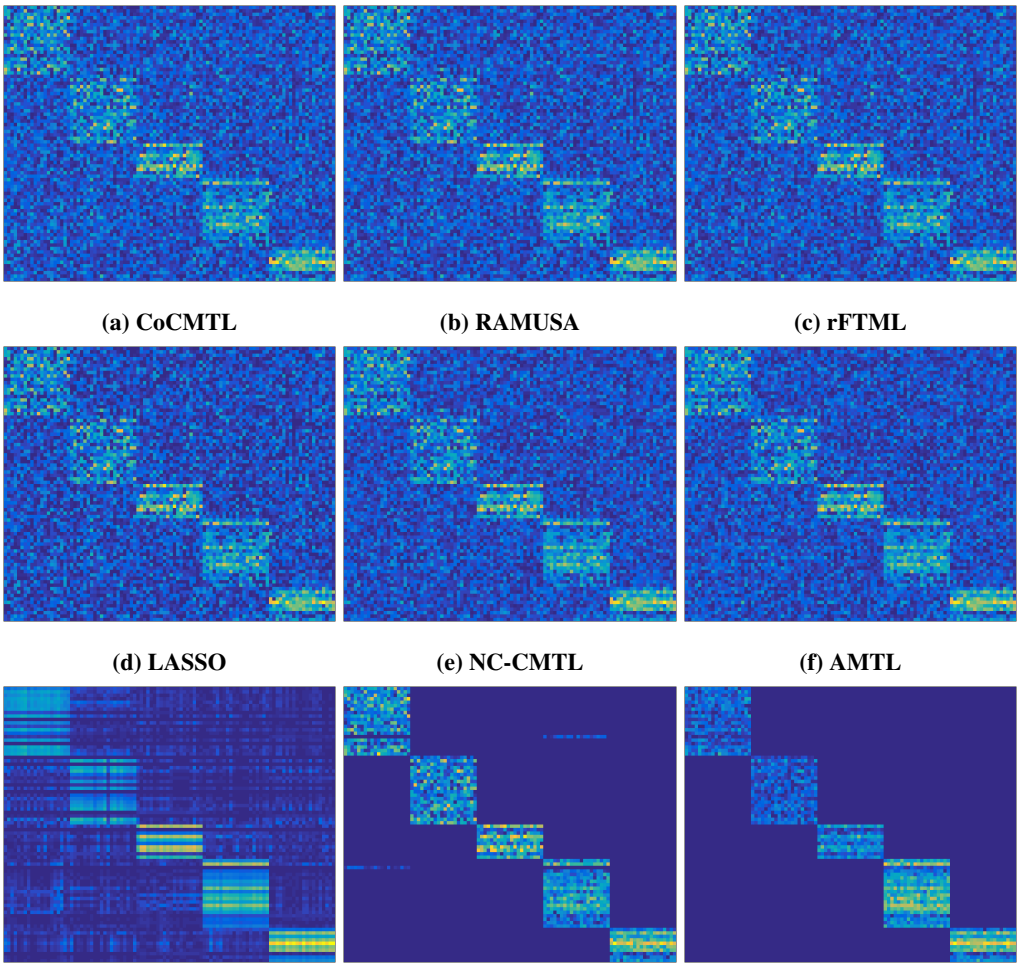


图 5.7 仿真数据集上的块对角结构恢复

Figure 5.7 Structural Recovery on Simulation Dataset

Analysis，PCA）过滤冗余输入特征。同时，为消除仅提供单类标签的极端用户产生的影响，手动移除为某些类给出少于8份标注的用户。

性能对比： Shoes数据集上15轮重复实验的平均结果如图5.8左侧所示，其中散点图代表由不同数据集划分取得的AUC性能，而柱状图展示15次重复实验的平均AUC性能。可以得到如下结论：1）针对Shoes数据集中所有属性，所提出的方法显著优于所有对比方法；相较于次优方法，*BR*、*CM*、*FA*、*FM*、*OP*、*OR*和*PT*的AUC评分分别提高了6.22、1.88、2.34、 2.38、3.66、 3.27及2.85。2）由于用户标注间存在显著的相关性，而低秩假设能够更有效地建模相关性，因而基于低秩约束的模型均优于基于稀疏约束的模型（LASSO和rMTFL）。3）由于通过不对称学习显式建模并消除负迁移的影响，AMTL优于所有其他基于低秩约束的模型。4）所提出方法优于其他学习特征-任务相关性的方法，证明

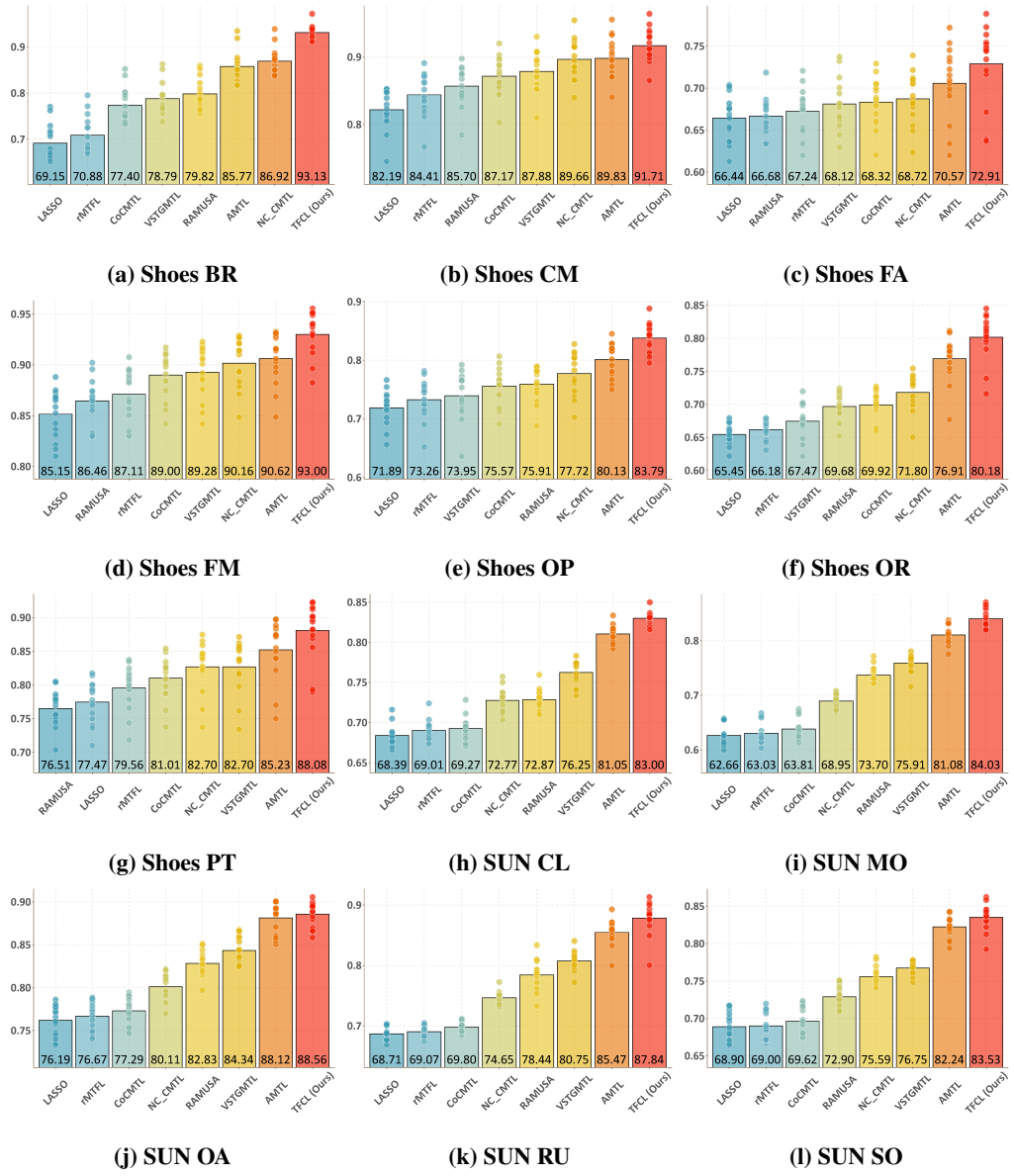


图 5.8 AUC指标对比图

Figure 5.8 Overall AUC comparisons

了TFCL框架的合理性。5）所提出方法在大多数属性上优于AMTL，一方面来自对用户标注行为合理的建模，另一方面来自更强负迁移抑制机制。

消融实验： 以下展示所提出方法包含的两个模块，即群体因素正则化算子和替代AUC损失，对模型性能的影响。(I) 具体而言，图5.9a-图5.9g分别展示了自原始模型移除上述两个模块后的实验结果，其中**Org**为原始模型性能，**w/o_AUC**为使用逐样本均方损失代替AUC损失对应的性能，而**w/o_G**代表移除所提出的协同分组因素对应的性能。根据实验结果，有以下结论：1）原模型优

于上述两种对比方法，说明两个模块的综合效应优于单个模块。2）大多数情况下，移除分组因素导致更显著的性能衰减，说明相较于AUC损失，群体效应对性能影响更大。(II) 由于CocMTL同样设计了一种协同分组正则化算子，因此进一步记录将所提出正则化算子替换为CoCMTL中对应项的实验结果。如图5.9a-图5.9g所示，所提出的正则化算子优于CoCMTL中的协同分组正则化算子。

细粒度对比： 图5.11a以属性*brown*为例，可视化所有方法用户测试AUC评分分布，从而实现细粒度对比。相较于对比方法，所提出的方法性能均值更高，而方差更小。相反，传统方法存在明显的长尾问题，其原因可能来自两方面：1）传统方法对困难任务更敏感。2）未能够有效地抑制负迁移。综上所述，所提出的方法确实促进高效的协同学习，从而提高了困难任务上的性能。

5.5.4.3 Sun数据集

预处理： 数据预处理与Shoes数据集的唯一区别在于使用Inception-V3(Szegedy等, 2016)网络抽取2048维特征向量作为输入特征。Shoes数据集中图片背景均为纯白色，而Sun数据集中图片背景则较为复杂，因此使用不同的特征提取器。

性能对比： Sun数据集上15轮重复实验的平均结果如图5.8所示，其中散点图代表由不同数据集划分取得的AUC性能，而柱状图展示15次重复实验的平均AUC性能。同Shoes数据集结果相似，有以下结论：1）所提出的方法在所有属性上均显著优于所有的对比方法，相较于次优模型在属性*CL*、*MO*、*OA*、*RU*和*SO*上AUC 性能分别提高了1.95、2.95、0.44、2.37和1.29。2）其他结论同Shoes数据集上2）-5）一致。

消融实验： 同Shoes数据集相似，图5.9h-Fig.5.9l和图5.10h-Fig.5.10l展示了Sun数据集上对应的消融结果。根据实验结果，有以下结论：1）原方法在除了SO的所有属性上均优于对比方法。2）相对于移除平方AUC损失，移除群体因素产生相对显著的性能衰减。3）所提出的正则化算子优于CoCMTL中的协同分组正则化算子。

细粒度对比： 图5.11a以属性*Open Area*为例，可视化所有方法用户测试AUC评分分布。如图中所示，所提出方法取得更为紧凑的评分分布，再一次说明所提出方法能够有效地抑制负迁移。

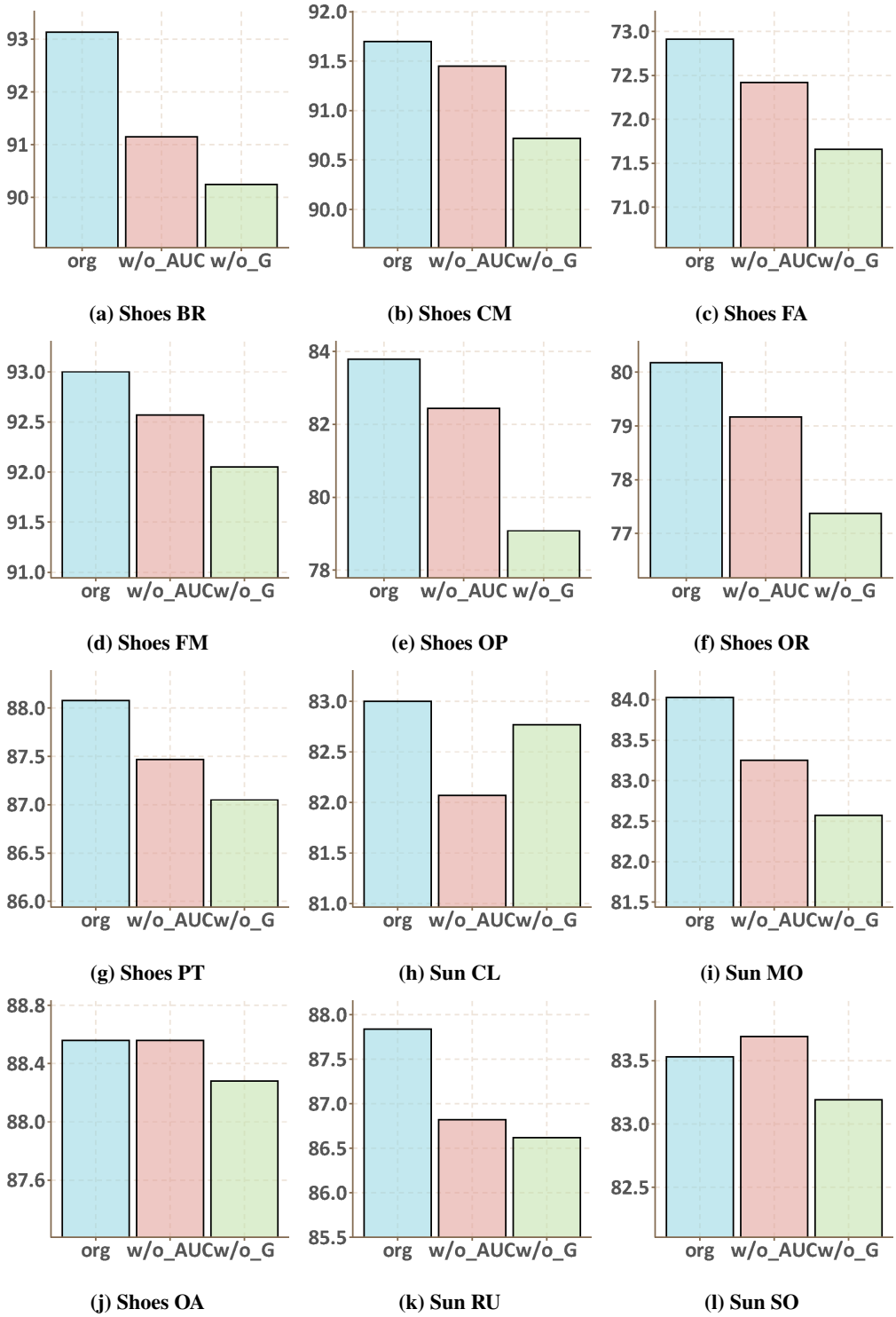


图 5.9 消融实验结果(I)

Figure 5.9 Ablation Results (I)[3]

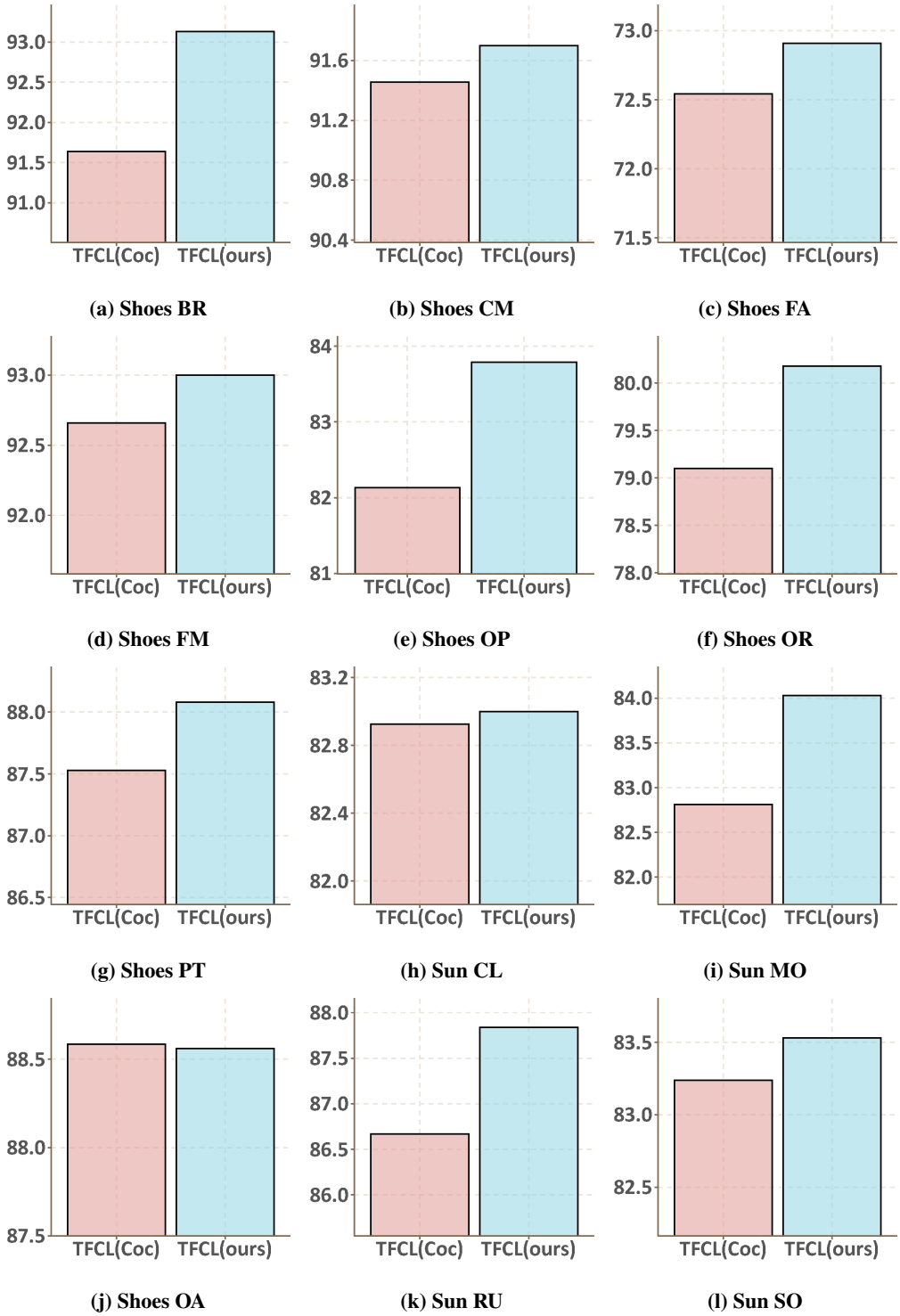
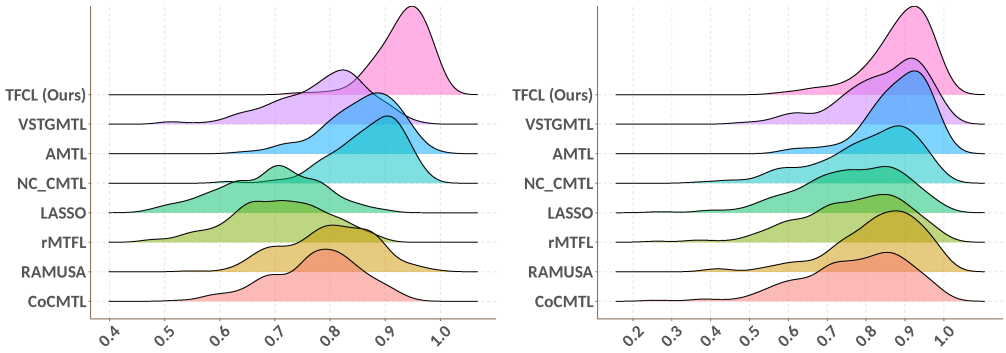


图 5.10 消融实验的结果示意图(II)

Figure 5.10 Ablation Results (II)[4]



(a) Shoes Brown　　**(b) SUN Open Area**

图 **5.11** 基于用户**AUC**分数分布的细粒度的比较

Figure 5.11 Fine-grained comparison based on User AUC Score Distributions

5.6 小结

本章提出一种名为TFCL的多任务学习方法，通过协同分组正则化算子同时在特征和任务层面抑制负迁移。同时，设计优化算法将原问题分解为两个迭代求解的凸子问题，并为其中一子问题提供闭式解，从而实现算法全局收敛性的理论证明。此外，由优化算法得到的解显示所提出方法与最优传输问题间存在紧密的联系，为抑制特征和任务负迁移提供了新思路。本章进一步通过多层模型分解机制将TFCL框架应用于个性化属性学习问题。为验证所提出方法的有效性，在一个仿真数据集和两个真实数据集开展系统性实验。仿真数据集上结果显示TFCL能够有效恢复正确的分组结构，而真实数据集上的结果则进一步验证所提出方法对于个性化属性预测问题的有效性。

第6章　总结与展望

本文工作围绕半监督、多分类、多任务场景下的AUC优化方法、理论及应用展开了系列研究，主要工作总结如下：

提出一种基于boosting的无先验半监督AUC优化模型集成方法，以解决半监督AUC优化的泛化性能瓶颈。首先，针对AUC设计高效的加速算法，将更新单个弱分类器的时间复杂度由平方降至线性。对所得算法进行了系统研究：在算法收敛速率方面，在一定假设下证明训练集误差随弱分类器个数增加以指数速率快速收敛。在算法泛化性能方面，首次提出半监督AUC优化的Rademacher复杂度，并提出一种广义最大值不等式，在此基础上首次给出了在模型集成意义下的半监督AUC优化泛化误差上界。

提出一种基于M度量的多分类AUC指标，以拓展AUC指标在多分类任务中的实用性。构建了多分类AUC的替代经验风险最小化问题，并展开了系统的理论分析。一致性方面，证明对于多分类AUC指标，指数损失、logit损失、平方损失、hinge损失等损失函数在特定假设下均具有一致性。泛化能力方面，首次系统分析了深度全连接网络及深度卷积网络上多分类AUC优化的Rademacher复杂度及覆盖数。最后，为hinge损失、平方损失以及指数损失的损失及梯度计算设计了高效的加速算法，大幅提升了多分类AUC优化的可拓展性。

聚焦于个性化属性学习问题，进一步将AUC优化应用于多任务学习中。从多任务学习视角出发，将每个用户的属性学习问题视为不同的任务，并将各用户模型参数分解为主流用户共识、用户群体聚类和个性化三级要素，以达到对用户共性和个性的同时建模。基于此，引入对AUC的直接优化构造得到目标函数，提出一种近端梯度下降优化方法进行求解，并针对AUC设计了一种加速计算方法，大幅提升计算效率。最后，系统分析了所提方法的收敛性和泛化能力，理论和实验结果均证明所提出方法的优越性。

在多任务AUC优化基础上，提出一种基于任务-特征协同的多任务学习框架，以抑制多任务学习中的负迁移问题。首先研究异构二分图的谱图论性质，提出一种块对角谱正则约束，通过促进任务-特征协同分组来保证组内知识共享，并阻断组间的知识迁移，实现负迁移抑制。针对所提非凸非光滑目标函数，

提出一种类近端梯度优化方法交替求解模型参数。同时，对该框架展开了系统的理论分析，证明所提出算法可保障所求非凸问题全局收敛性质并有效恢复其块对角结构。在收敛性质方面，首次证明该类优化问题（截断特征根求和最小化）的参数序列全局收敛性质。而在结构恢复方面，首次给出在非渐近意义下无假阳性结构恢复与准确结构恢复所需的超参数确定性选取条件及基本假设。最后，在个性化属性学习任务上验证了所提出框架的有效性。

未来工作展望

本文虽然围绕复杂场景下的AUC优化问题进行了一定程度的研究并取得了一定成果，但研究过程中仍存在部分问题有待进一步解决，下面尝试给出一些未来研究中值得关注的开放性问题：

(A) 多分类AUC优化方面，本文所提出的加速算法可有效减小单次迭代的复杂度，并在实验中取得了较为显著的加速结果。然而，本文工作并未涉及如何通过改进优化算法来减少模型达到固定精度所需的迭代次数。因此未来工作中，还需进一步针对多分类AUC优化问题本身设计收敛率更高的优化算法。

(B) 多任务AUC优化方面，本文所提出的基于负迁移机制的模型，特征选择方法可进一步应用于基于神经网络的深度学习框架中，辅助神经网络进行结构选择。如何在复杂的神经网络中形式化特征、模块、输出之间的关联矩阵，如何构建可拓展的优化算法无疑都是极具挑战及意义的研究方向。另一方面，在多分类问题中，类与类之间同样可能存在复杂的迁移关系，如何让负迁移抑制服务于多分类AUC优化问题也是十分有意义的未来研究方向。

(C) 除本文涉及的半监督、多分类、多任务的复杂场景外，AUC指标本身也存在一定的局限性。例如，引言中已提及，AUC指标刻画了所有(TPR, FPR)取值下模型的平均性能，而有效的模型性能一般多集中于高TPR及低FPR，因此如何优化ROC的局部面积也是一个关键的挑战。在现有的相关工作主要采用割平面方法求解优化问题，难以适用端到端的训练模型，如何构建高效的ROC局部面积优化方法也是未来需要解决的一个关键问题。

(D) 在应用层面，可将AUC优化算法进一步拓展至其他更为复杂的应用场景中。例如，在知识图谱的复杂数据结构上应用AUC优化方法完成长尾链路预测等关键任务；在无监督/自监督场景下通过AUC优化方法进行长尾预测等。

附录 A　第2章的算法及证明补充

A.1　弱分类器的学习

给定已排序的输入特征集合$\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_d\}$（$\boldsymbol{X}^i$为所有训练样本第i为特征构成的集合），及已排序的阈值候选集$\boldsymbol{T} = \{\boldsymbol{T}^1, \cdots, \boldsymbol{T}^d\}$（$\boldsymbol{T}^i$第i维特征候选阈值构成的集合），通过如下搜素方式完成最佳参数的选择。

算法 8 基于决策树桩的弱分类器

Require: 模型输入特征 $\boldsymbol{X} = \{\boldsymbol{X}^1, \cdots, \boldsymbol{X}^d\}$，其中$\boldsymbol{X}^i$为所有训练样本第i维特征数值的集合，令$X_j^i$表示第$j$个样本的第$i$维特征数值。进入算法前，预先对每个$\boldsymbol{X}^i$元素进行降序排序，使得：$X_1^i \geq X_2^i \cdots \geq X_N^i$。

输入: 阈值备选集$\boldsymbol{T} = \{\boldsymbol{T}^1, \cdots, \boldsymbol{T}^d\}$。其中$\boldsymbol{T}^i$为第i维特的阈值备选集集合，令$T_j^i$表示第$i$个特征的第$j$个阈值备选值，同样预先将其排序并使$T_1^i \geq T_2^i \cdots \geq T_N^i$。

输出: 最佳特征维度 $i^\star$, 最佳阈值 $\theta_{i^\star}^\star$, Δ^t

$RE_0 \leftarrow -\infty$

for $i \leftarrow 1 : d$ **do**

　$last \leftarrow -1$

　$L \leftarrow 0$

　for $j \leftarrow 1 : |T_i|$ **do**

　　$offset \leftarrow last + 1$

　　for $k \leftarrow (last + 1) : N$ **do**

　　$t \leftarrow T_j^i$

　　if $X_k^i > t$ **then**

　　$L+ = g_{t_i}$

　　$last \leftarrow k$

　　else

　　break

　　end if

　　end for

　　$RE_1 \leftarrow KL\left(\frac{1+\rho}{2} || \frac{1+L}{2}\right)$

　　if $RE_1 > RE_0$ **then**

　　$\Delta_t \leftarrow L$

　　$i\star \leftarrow i$

　　$\theta_{i^\star}^\star \leftarrow t$

　　$RE_0 \leftarrow RE_1$

　　end if

　end for

end for

A.2 引理 2.1 证明

证明. 展开算法1中**Step 4**的迭代规则，有

$$D^{T+1}(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)=\frac{D^0(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)\exp\left(f(\tilde{\boldsymbol{x}}_j)-f(\boldsymbol{x}_i)\right)}{\prod_{t=1}^{T}\tilde{Z}^t}, \tag{A.1}$$

$$D^{T+1}(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})=\frac{D^0(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})\exp\left(f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_k^{(n)})\right)}{\prod_{t=1}^{T}\tilde{Z}^t}, \tag{A.2}$$

$$D^{T+1}(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})=\frac{D^0(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})\exp\left(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_k^{(n)})\right)}{\prod_{t=1}^{T}\tilde{Z}^t}, \tag{A.3}$$

将式 (A.1)，式 (A.2)和式 (A.3)代入式 (2.21)得：

$$\begin{aligned}R_{OP4}=\prod_{t=1}^{T}\exp(-\rho\cdot\alpha^t)\cdot\tilde{Z}^t\Bigg(&\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}D^{T+1}(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)\\&+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})+\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})\Bigg)\end{aligned} \tag{A.4}$$

由式 (2.17)-式 (2.19)，式 (2.24)-式 (2.30)，式 (2.31)-式 (2.36)，易得样例对权重D^{T+1}在算法中已被归一化，即：

$$\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}D^{T+1}(\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_j)+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_j^{(u)},\boldsymbol{x}_k^{(n)})+\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}D^{T+1}(\boldsymbol{x}_i^{(p)},\boldsymbol{x}_k^{(n)})=1$$

由式 (A.4)，式 (A.2) 此引理得证。 □

A.3 引理 2.2 证明

证明. 首先对式 (2.38)中的$C(\alpha^t)$进行化简，有：

$$\begin{aligned}\exp\left(\rho\alpha^t\right)\cdot C\left(\alpha^t\right)=&\frac{1+\Delta^t}{2}\exp\left((\rho-1)\cdot\alpha^t\right)\\&+\frac{1-\Delta^t}{2}\exp\left((\rho+1)\cdot\alpha^t\right)\end{aligned} \tag{A.5}$$

将式 (2.43) 代入式 (A.5)，将$\exp\left(\rho\alpha^t\right)\cdot C\left(\alpha^t\right)$ 重写为：

$$\begin{aligned}&\underbrace{\frac{1-\Delta^t}{2}\exp\left[\frac{\rho+1}{2}\log\left(\frac{(1+\Delta^t)\cdot(1-\rho)}{(1-\Delta^t)\cdot(1+\rho)}\right)\right]}_{(I)}\\&+\underbrace{\frac{1+\Delta^t}{2}\exp\left[\frac{\rho-1}{2}\log\left(\frac{(1+\Delta^t)\cdot(1-\rho)}{(1-\Delta^t)\cdot(1+\rho)}\right)\right]}_{(II)}\end{aligned} \tag{A.6}$$

对于(I)，有：

$$\begin{aligned}(I) &= \frac{1}{2}\exp\left[-KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)+\log\left(\frac{1-\rho}{1-\Delta^t}\right)+\log(1-\Delta^t)\right] \\ &= \frac{1-\rho}{2}\cdot\exp\left[-KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)\right]\end{aligned} \tag{A.7}$$

对于(II)，有：

$$\begin{aligned}(II) &= \frac{1}{2}\exp\left[-KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)+\log\left(\frac{1+\rho}{1+\Delta^t}\right)+\log(1+\Delta^t)\right] \\ &= \frac{1+\rho}{2}\cdot\exp\left[-KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)\right]\end{aligned} \tag{A.8}$$

联立式 (A.6)、式 (A.7)及式 (A.8)，有：

$$\begin{aligned}\exp\left(\rho\alpha^t\right)C\left(\alpha^t\right) &= \left(\frac{1+\rho}{2}+\frac{1-\rho}{2}\right)\cdot\exp\left[-KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)\right] \\ &= \exp\left[-KL\left(\frac{1+\rho}{2}||\frac{1+\Delta^t}{2}\right)\right]\end{aligned}$$

由此引理得证。 □

A.4　定理2的证明

A.4.1　预备引理

定义 A.1 (有限差分性质(Bounded Difference Property)). 给定一组独立随机变量$X_1,\cdots,X_n$，及其定义域$\mathbb{X}$，对于函数$f(X_1,X_2,\cdots,X_n)$若存在非负常量$c_1,c_2,\cdots,c_n$使得：

$$\sup_{x_1,x_2,\cdots,x_n,x_i'}\left|f(x_1,\cdots,x_n)-f(x_1,\cdots,x_{i-1},x_i',\cdots,x_n)\right|\leq c_i, \quad \forall 1\leq i\leq n \tag{A.9}$$

则称函数f满足有限差分性质。

对于作用于随机变量上且满足该性质的函数，有以下不等式：

引理 A.1 (有限差分不等式（Bounded Difference Inequality). (见文献(Boucheron 等, 2013, 命题6.1及定理6.2)) 设$X_1,\cdots,X_n$，且$X_i\in\mathbb{X}$ 为一组独立随机变量，令$Z=f(X_1,\cdots,X_n)$，若f满足有限差分性质，且对应常数为$c_1,c_2,\cdots,c_n$，则有：

$$\log\mathbb{E}\left[\exp\left(\lambda(Z-\mathbb{E}[Z])\right)\right]\leq\frac{\lambda^2 v}{2}, \tag{A.10}$$

对于任意$\lambda>0$成立，其中：

$$v=\frac{1}{4}\sum_{i=1}^{n}c_i^2 \tag{A.11}$$

引理 A.2 (最大值不等式（Maximal Inequality）). (见文献(Boucheron 等, 2013, 2.5节)) 令 $Z_1, \cdots, Z_n$ 为一组实数值随机变量且存在 $v > 0$使任意$i = 1, 2, \cdots, n$，有$\log\left(\mathbb{E}\left[\exp\left(\lambda Z_i\right)\right]\right) \leq \frac{\lambda^2 v}{2}$, 则有:

$$\mathbb{E}\left[\max_{i=1,2,\cdots,n} Z_i\right] \leq \sqrt{2v \log n}.$$

引理 A.3 (Mcdiarmid 不等式). （见文献(McDiarmid, 1998)）令 $X_1, \cdots, X_m$ 为一组取值于集成$\mathcal{X}$内的独立随机变量，令 $f : \mathcal{X} \to \mathbb{R}$ 为满足:

$$\sup_{\boldsymbol{x}, \boldsymbol{x}'} |f(x_1, \cdots, x_i, \cdots, x_m) - f(x_1, \cdots, x_i' \cdots, x_,)| \leq c_i,$$

其中 $\boldsymbol{x} \neq \boldsymbol{x}'$，的函数，则对于任意$\epsilon > 0$有：

$$\mathbb{P}[\mathbb{E}(f) - f \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

$$\mathbb{P}[f - \mathbb{E}(f) \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

引理 A.4 (Talagrand压缩引理). 令$\ell_1, \cdots, \ell_l$为一组ϕ-Lipshiptz 连续函数，$\sigma_1, \cdots, \sigma_m$为相互独立的一组Rademacher随机变量，则：

$$\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{H}} \sum_{i=1}^m \sigma_i \cdot (\ell_i \circ f)(x)\right] \leq \frac{\phi}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{H}} \sum_{i=1}^m \sigma_i \cdot f(x)\right]$$

引理 A.5. 见文献(Mohri 等, 2018, 引理7.4) 给定$\mathcal{H}$为由函数$f : \mathcal{X} \to \mathbb{R}$构成的函数集，记 $co\left(\mathcal{H}\right)$为:

$$co\left(\mathcal{H}\right) = \left\{f : \sum_{i=1}^T \alpha^t h^t, \sum_{i=1}^T \alpha^t = 1, \alpha^t \geq 0, \quad h^t \in \mathcal{H}\right\}$$

$$\frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in co(\mathcal{H})} \sum_{i=1}^m \sigma_i \cdot f(x)\right] = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{H}} \sum_{i=1}^m \sigma_i \cdot f(x)\right]$$

A.4.2 本文提出的引理

引理 A.6 (广义最大值不等式). 给定实数值随机变量$\{M_i^{(k)}\}_{1 \leq i \leq N,\ 1 \leq k \leq K}$，若满足以下条件：

1. 对于任意$k_1 \neq k_2$， $M_{i_1}^{(k_1)}$ 与 $M_{i_2}^{(k_2)}$，相互独立。
2. $\forall i,\ k$, $\mathbb{E}[M_i^{(k)}] = 0$，且 $\log\left(\mathbb{E}\left[\exp\left(\lambda M_i^{(k)}\right)\right]\right) \leq \dfrac{\lambda^2 v_k}{2}$

有：

$$\mathbb{E}\left(\sum_{k=1}^{K}\max_{i} M_i^k\right) \leq (2\log N \cdot \sum_k v_k)^{1/2}$$

证明. 见附录A.4.4。 □

引理 A.7 (PNU-AUC的对称化技术). 将数据集$\mathcal{S}$及$\mathcal{S}'$的类标集固定为$\boldsymbol{Y}, \boldsymbol{Y}'$, 则对任意假设集$\mathcal{H}$及损失函数$\ell$，以下结论成立：

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right)\right] \leq 4\mathfrak{R}_{PNU}(\ell\circ\mathcal{H}),$$

证明. 见附录A.4.6。 □

引理 A.8.

$$\hat{\mathfrak{R}}_{PNU}(\ell_\rho \circ co\,(\mathcal{H}_{DS})) \leq 2\cdot\left(2(\log d + \log K)\cdot\rho_\gamma(\boldsymbol{Y})\right)^{1/2}$$

其中

$$\rho_\gamma(\boldsymbol{Y}) = \left(\frac{1}{n_p} + \frac{1}{n_u} + \frac{1}{n_n}\right)$$

证明. 见附录A.4.6。 □

A.4.3　定理2证明

证明.

步骤1：固定所有训练样本类别$\boldsymbol{Y}$，由引理 A.3 中的第二式构造

$$\sup_{f\in co(\mathcal{H}_{DS})}\left[\mathop{\mathbb{E}}_{\mathcal{S}'}\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right]$$

的大概率上界，且在上界中引入PNU-Rademacher复杂度。

由引理A.7, 有:

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in co(\mathcal{H}_{DS})}\left(\hat{R}_{\mathcal{S}}(f) - \hat{R}_{\mathcal{S}'}(f)\right)\right] \leq 4\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell\circ\mathcal{H}). \tag{A.12}$$

给定训练样本 $\mathcal{S}$，定义 $\mathcal{S}_o = (\mathcal{S}\setminus\{(\boldsymbol{x}_o, y_o)\})\cup\{(\boldsymbol{x}'_o, y_o)\}$. 考察

$$\sup_{f\in co(\mathcal{H}_{DS})}\left[\mathop{\mathbb{E}}_{\mathcal{S}'}\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right]$$

相对于训练样本的有限差分性质系数。若当前 $y_o = 1$，有:

$$\begin{aligned}
c_o &= \sup_{\boldsymbol{x}_o^{(p)}, \tilde{\boldsymbol{x}}_o^{(p)}} \left| \sup_{f \in co(\mathcal{H}_{DS})} (\mathop{\mathbb{E}}_{\mathcal{S}} \hat{R}_{\mathcal{S}}(f) - \hat{R}_{\mathcal{S}}(f)) - \sup_{f \in co(\mathcal{H}_{DS})} (\mathbb{E}_{\mathcal{S}_i} \hat{R}_{\mathcal{S}_i}(f) - \hat{R}_{\mathcal{S}_i}(f)) \right| \\
&\leq \sup_{\boldsymbol{x}_o^{(p)}, \tilde{\boldsymbol{x}}_o^{(p)}} \sup_{f \in co(\mathcal{H}_{DS})} \left| \hat{R}_{\mathcal{S}}(f) - \hat{R}_{\mathcal{S}_i}(f) \right| \\
&= \sup_{\boldsymbol{x}_o^{(p)}, \tilde{\boldsymbol{x}}_o^{(p)}} \sup_{f \in co(\mathcal{H}_{DS})} \left[\frac{\gamma}{n_p n_n} \sum_{k=1}^{n_n} |\ell_\rho(f, \boldsymbol{x}_o^{(p)}, \boldsymbol{x}_k^{(n)}) - \ell_\rho(f^{(i)}, \tilde{\boldsymbol{x}}_o^{(p)}, \boldsymbol{x}_k^{(n)})| \right. \\
&\quad \left. + \frac{1-\gamma}{2n_p n_u} \sum_{j=1}^{n_u} |\ell_\rho(f, \boldsymbol{x}_o^{(p)}, \boldsymbol{x}_j^{(u)}) - \ell_\rho(f, \tilde{\boldsymbol{x}}_o^{(p)}, \boldsymbol{x}_j^{(u)})| \right] \\
&\overset{(1)}{\leq} (\gamma + \frac{1-\gamma}{2}) \frac{1}{\rho \cdot n_p} \\
&\leq \frac{1}{\rho \cdot n_p}
\end{aligned}$$

同理，当$y_o = -1$有$c_o \leq \frac{1}{\rho \cdot n_n}$；当$y_o = 0$有$c_o \leq \frac{1}{\rho \cdot n_u}$。此时令

$$\begin{aligned}
v =& \frac{1}{\rho} \sum_{o=1}^{N} c_o^2 = \frac{1}{\rho} \cdot \max\left\{\gamma^2, \frac{(1-\gamma)^2}{4}\right\} \left(\frac{1}{n_p} + \frac{1}{n_u} + \frac{1}{n_n}\right) \\
\leq& \frac{1}{\rho} \cdot \left(\frac{1}{n_p} + \frac{1}{n_u} + \frac{1}{n_n}\right)
\end{aligned}$$

对$\sup_{f \in co(\mathcal{H}_{DS})} \left[\mathbb{E}_{\mathcal{S}'} \hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right]$运用引理 A.3 第二式，可知，下式至少以$1 - \frac{\delta}{2}$概率成立：

$$\begin{aligned}
&\sup_{f \in co(\mathcal{H}_{DS})} \left[\mathop{\mathbb{E}}_{\mathcal{S}'} \hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right] \\
&\leq \mathop{\mathbb{E}}_{\mathcal{S}} \sup_{f \in co(\mathcal{H}_{DS})} \left[\mathop{\mathbb{E}}_{\mathcal{S}'} \hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right] \\
&\quad + \frac{1}{\rho} \left(\log(\frac{2}{\delta}) \cdot (\frac{1}{2n_p} + \frac{1}{2n_u} + \frac{1}{2n_n})\right)^{1/2}
\end{aligned} \tag{A.13}$$

为引入PNU-Rademacher复杂度，构造

$$\mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{f \in co(\mathcal{H}_{DS})} \left(\mathop{\mathbb{E}}_{\mathcal{S}'} \hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f) \right) \right]$$

的上界，根据Jensen不等式:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{f \in co(\mathcal{H}_{DS})} \left(\mathop{\mathbb{E}}_{\mathcal{S}'} \hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f) \right) \right] \\
=& \mathop{\mathbb{E}}_{\mathcal{S}} \sup_{f \in co(\mathcal{H}_{DS})} \mathop{\mathbb{E}}_{\mathcal{S}'} \left[\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f) \right] \\
\leq& \mathop{\mathbb{E}}_{\mathcal{S}, \mathcal{S}'} \left[\sup_{f \in co(\mathcal{H}_{DS})} \left(\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f) \right) \right]
\end{aligned}$$

因此有对于任意$f \in co\left(\mathcal{H}_{DS}\right)$，下式至少以概率$1-\frac{\delta}{2}$

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}} \hat{R}_{\mathcal{S}}(f) \leq & \hat{R}_{\mathcal{S}}(f)+4 \mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell \circ \mathcal{H}) \Gamma \\ & +\frac{1}{\rho}\left(\log (\frac{2}{\delta}) \cdot (\frac{1}{2 n_p}+\frac{1}{2 n_u}+\frac{1}{2 n_n})\right)^{1 / 2}\end{aligned} \tag{A.14}$$

至此步骤1证毕。

步骤2：固定样例类别，对$\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell \circ \mathcal{H})$运用引理 A.3 第一式，在大概率上界中由$\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H})$替换$\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell \circ \mathcal{H})$，并由引理 A.8将$\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H})$化简。

通过再次考察$\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H})$相对于训练样本的有限差分性质，可得下式至少以$1-\frac{\delta}{2}$概率成立：

$$\begin{aligned}\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell \circ \mathcal{H}) \leq & \hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H}) \\ & +\frac{1}{\rho}\left(\log (\frac{2}{\delta}) \cdot (\frac{1}{2 n_p}+\frac{1}{2 n_u}+\frac{1}{2 n_n})\right)^{1 / 2}\end{aligned} \tag{A.15}$$

联立 式 (A.14)) 及式 (A.15))以及概率的次可加性，下式以 $1-\delta$概率成立：

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}} \hat{R}_{\mathcal{S}}(f) \leq & \hat{R}_{\mathcal{S}}(f)+4 \hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H}) \\ & +\frac{1}{\rho}\left(\log (\frac{2}{\delta}) \cdot (\frac{1}{2 n_p}+\frac{1}{2 n_u}+\frac{1}{2 n_n})\right)^{1 / 2}\end{aligned} \tag{A.16}$$

代入引理 A.4.6 ，有：

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}} \hat{R}_{\mathcal{S}}(f) \leq & \hat{R}_{\mathcal{S}}(f)+\frac{8 \sqrt{2}}{\rho}\left((\log d+\log K) \cdot \chi(\boldsymbol{Y})\right)^{1 / 2} \\ & +\frac{1}{\rho}\left(\log (\frac{2}{\delta}) \cdot \chi(\boldsymbol{Y})\right)^{1 / 2}\end{aligned}$$

至此步骤2证毕

步骤3：解除对$\boldsymbol{Y}$的固定。

推导过程同工作(Agarwal 等, 2005)中定理 8，可证明式 (A.16) 相对于整个$\mathcal{S}$ 仍可以至少$1-\delta$的概率成立。

步骤4：完成主结论证明。

由式 (2.69) 及简单推导化简可推知：$\hat{R}_{\mathcal{S}}(f) \leq r_\rho$以及$\mathbb{E}_{\mathcal{S}}\left[\hat{R}_{0-1, S}^{PNU}\right] \leq \mathbb{E}_{\mathcal{S}}\left[\hat{R}_{\mathcal{S}}(f)\right]$，此时有，下式以至少$1-\delta$概率成立：

$$\begin{aligned}R_{0-1}^{PNU}(f) \leq & r_\rho+\frac{8 \sqrt{2}}{\rho}\left((\log d+\log K) \cdot \chi(\boldsymbol{Y})\right)^{1 / 2} \\ & +\frac{1}{\rho}\left(\log (\frac{2}{\delta}) \cdot \chi(\boldsymbol{Y})\right)^{1 / 2}\end{aligned} \tag{A.17}$$

综合式 (2.3)及式 (2.4)，此时有：

$$\begin{aligned} R_{0-1}^{PNU}(f) &= \gamma \cdot R_{0-1}^{PN} + \frac{1-\gamma}{2}(R_{0-1}^{PU} + R_{0-1}^{UN}) \\ &= \gamma \cdot R_{0-1}^{PN} + \frac{1-\gamma}{2}(R_{0-1}^{PN} + \frac{1}{2}) \\ &\geq \frac{1+\gamma}{2} R_{0-1}^{PN} \end{aligned} \tag{A.18}$$

综合式 (A.17) 及 式 (A.18)，定理得证。 □

A.4.4 引理 A.6 证明

证明.

$$\begin{aligned} & \exp\left(\mathbb{E}\left(\lambda \sum_{k=1}^{K} \max_i M_i^k\right)\right) \\ \overset{(1)}{\leq}\ & \mathbb{E}\left(\prod_{k=1}^{K} \exp\left(\lambda \max_i M_i^k\right)\right) \\ \overset{(2)}{=}\ & \prod_{k=1}^{K} \mathbb{E}\left(\exp\left(\lambda \max_i M_i^k\right)\right) \\ \overset{(3)}{=}\ & \prod_{k=1}^{K} \mathbb{E}\left(\max_i \exp\left(\lambda M_i^k\right)\right) \\ \leq\ & \prod_{k=1}^{K} \mathbb{E}\left(\sum_{i=1}^{N} \exp\left(\lambda M_i^k\right)\right) \\ \overset{(4)}{\leq}\ & N \exp\left(\frac{\lambda^2 \sum_{k=1}^{K} v_k}{2}\right) \end{aligned}$$

其中(1)有Jensen不等式获得、(2)由本引理假设条件1获得、(3)由指数函数$\exp(x)$的严格单增性获得、(4)由本引理假设条件2获得。由上式进一步获得：

$$\mathbb{E}\left(\sum_{k=1}^{K} \max_i M_i^k\right) \leq \frac{\log\left(N \exp\left(\frac{\lambda^2 \sum_{k=1}^{K} v_k}{2}\right)\right)}{\lambda}$$

对上式求最大值，引理得证。 □

A.4.5　引理 A.7 证明

证明. 此处记 $\ell(f,\boldsymbol{x_1},\boldsymbol{x_2})=\ell(f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2))$ 定义

$$
\begin{aligned}
Q_\sigma^{p,n,i,k} =& \frac{\sigma_i^{(p)}+\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{(\boldsymbol{p})},\tilde{\boldsymbol{x}}_{\boldsymbol{k}}^{(\boldsymbol{n})})+\frac{\sigma_i^{(p)}-\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{(\boldsymbol{p})},\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{n})}) \\
&-\frac{\sigma_i^{(p)}-\sigma_k^{(n)}}{2}\ell(f,\boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{p})},\tilde{\boldsymbol{x}}_{\boldsymbol{k}}^{(\boldsymbol{n})})-\frac{\sigma_i^{(p)}+\sigma_k^{(n)}}{2}\ell(f,\boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{p})},\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{n})}), \\
Q_\sigma^{p,u,i,j} =& \frac{\sigma_i^{(p)}+\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{(\boldsymbol{p})},\tilde{\boldsymbol{x}}_{\boldsymbol{j}}^{(\boldsymbol{u})})+\frac{\sigma_i^{(p)}-\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{(\boldsymbol{p})},\boldsymbol{x}_{\boldsymbol{j}}^{(\boldsymbol{u})}) \\
&-\frac{\sigma_i^{(p)}-\sigma_k^{(n)}}{2}\ell(f,\boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{p})},\tilde{\boldsymbol{x}}_{\boldsymbol{j}}^{(\boldsymbol{u})})-\frac{\sigma_i^{(p)}+\sigma_j^{(u)}}{2}\ell(f,\boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{p})},\boldsymbol{x}_{\boldsymbol{j}}^{(\boldsymbol{u})}), \\
Q_\sigma^{u,n,j,k} =& \frac{\sigma_j^{(u)}+\sigma_k^{(n)}}{2}\ell(f,\tilde{\boldsymbol{x}}_{\boldsymbol{j}}^{(\boldsymbol{u})},\tilde{\boldsymbol{x}}_{\boldsymbol{k}}^{(\boldsymbol{n})})+\frac{\sigma_i^{(p)}-\sigma_j^{(u)}}{2}\ell(f,\tilde{\boldsymbol{x}}_{\boldsymbol{j}}^{(\boldsymbol{u})},\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{n})}) \\
&-\frac{\sigma_j^{(u)}-\sigma_k^{(n)}}{2}\ell(f,\boldsymbol{x}_{\boldsymbol{j}}^{(\boldsymbol{u})},\tilde{\boldsymbol{x}}_{\boldsymbol{k}}^{(\boldsymbol{n})})-\frac{\sigma_j^{(u)}+\sigma_k^{(n)}}{2}\ell(f,\boldsymbol{x}_{\boldsymbol{j}}^{(\boldsymbol{u})},\boldsymbol{x}_{\boldsymbol{k}}^{(\boldsymbol{n})}),
\end{aligned}
\tag{A.19}
$$

首先证明：

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}_{DS}}\left(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f)\right)\right] \\
&=\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\gamma}{n_pn_n}\cdot Q_\sigma^{p,n,i,k}+\right. \\
&\left.\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{1-\gamma}{2n_pn_u}\cdot Q_\sigma^{p,u,i,j}+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{1-\gamma}{2n_un_n}\cdot Q_\sigma^{u,n,j,k}\right].
\end{aligned}
\tag{A.20}
$$

给定样本集

$$
\begin{aligned}
\mathcal{S}&=\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}\cup\{\boldsymbol{x}_j^{(u)}\}_{j=1}^{n_u}\cup\{\boldsymbol{x}_k^{(n)}\}_{k=1}^{n_n} \\
\mathcal{S}'&=\{\tilde{\boldsymbol{x}}_i^{(p)}\}_{i=1}^{n_p}\cup\{\tilde{\boldsymbol{x}}_j^{(u)}\}_{j=1}^{n_u}\cup\{\tilde{\boldsymbol{x}}_k^{(n)}\}_{k=1}^{n_n}
\end{aligned}
$$

考虑$\mathcal{S},\mathcal{S}'$中的样例由独立采样获得，因此交换数据集中的相同类标的样例不影响数据分布。考虑该数据交换过程，记交换数据集 $\mathcal{S}$ 及 $\mathcal{S}'$ 任意一个或多个对应位置的样例，也即交换任意一个或多个$\boldsymbol{x}_i^{(p)}$与$\tilde{\boldsymbol{x}}_i^{(p)}$，$\boldsymbol{x}_j^{(u)}$ 与 $\tilde{\boldsymbol{x}}_j^{(u)}$，或 $\boldsymbol{x}_k^{(n)}$与$\tilde{\boldsymbol{x}}_k^{(n)}$所得的两个新数据集分别为$\tilde{S}$及$\tilde{S}'$，易得如下等式关系：

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}^{PNU}(f)-\hat{R}_{\mathcal{S}}^{PNU}(f)\right)\right] \\
&=\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\widetilde{\mathcal{S}}'}^{PNU}(f)-\hat{R}_{\widetilde{\mathcal{S}}}^{PNU}(f)\right)\right]
\end{aligned}
\tag{A.21}
$$

受上式启发，为证明式 (A.20)，首先证明对于任意独立同分布 Rademacher 随机变量序列 $\sigma=\{\sigma_i^{(p)}\}_i\cup\{\sigma_j^{(u)}\}_j\cup\{\sigma_k^{(n)}\}_k$ 以及任意$\mathcal{S}$，$\mathcal{S}'$，存在交换后产生的数

据集$\tilde{S}^{\sigma}$ 及 $\tilde{S}'^{\sigma}$ 使得：

$$\begin{aligned}\sup_{f\in\mathcal{H}}\Bigg[&\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\gamma}{n_p n_n}\cdot Q_{\sigma}^{p,n,i,k}+\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{1-\gamma}{2n_p n_u}\cdot Q_{\sigma}^{p,u,i,j}\\&+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{1-\gamma}{2n_u n_n}\cdot Q_{\sigma}^{u,n,j,k}\Bigg]\\=\sup_{f\in\mathcal{H}}&\left[\hat{R}_{\tilde{\mathcal{S}}^{\sigma\prime}}(f)-\hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f).\right]\end{aligned}\tag{A.22}$$

下面通过数学归纳法证明式 (A.22)

步骤1：考虑平凡情况：$n_p=1,n_u=1,n_n=1$ $\mathcal{S}=\left\{\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(n)},\boldsymbol{x}_1^{(u)}\right\}$ $\mathcal{S}'=\left\{\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(u)},\tilde{\boldsymbol{x}}_1^{(n)}\right\},\sigma=(\sigma_1^{(p)},\sigma_1^{(u)},\sigma_1^{(n)})$。下面通过分类讨论证明式 (A.22) 成立：

情况1-1 $\sigma_1^{(p)}=1,\sigma_1^{(u)}=1,\sigma_1^{(n)}=1$，：

$$\begin{aligned}Q_{\sigma}^{p,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(n)}),\\Q_{\sigma}^{p,u,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(u)})-\ell(f,\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(u)}),\\Q_{\sigma}^{u,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(u)},\tilde{\boldsymbol{x}}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(u)},\boldsymbol{x}_1^{(n)}),\end{aligned}\tag{A.23}$$

比较Rademacher的基本定义，若令 $\tilde{\mathcal{S}}_{\sigma}=\mathcal{S}$，$\tilde{\mathcal{S}'}_{\sigma}=\mathcal{S}'$，则式 (A.22)成立

情况1-2 $\sigma_1^{(p)}=-1,\sigma_1^{(u)}=-1,\sigma_1^{(n)}=-1$。通过与情况1类似的推导方法，可知，若令 $\tilde{\mathcal{S}}_{\sigma}=\mathcal{S}'$，$\tilde{\mathcal{S}'}_{\sigma}=\mathcal{S}$，则式 (A.22)成立。

情况1-3 $\sigma_1^{(p)}=1,\sigma_1^{(u)}=1,\sigma_1^{(n)}=-1$，此时有：

$$\begin{aligned}Q_{\sigma}^{p,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\boldsymbol{x}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(n)}),\\Q_{\sigma}^{p,u,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(p)},\tilde{\boldsymbol{x}}_1^{(u)})-\ell(f,\boldsymbol{x}_1^{(p)},\boldsymbol{x}_1^{(u)}),\\Q_{\sigma}^{u,n,1,1}&=\ell(f,\tilde{\boldsymbol{x}}_1^{(u)},\boldsymbol{x}_1^{(n)})-\ell(f,\boldsymbol{x}_1^{(u)},\tilde{\boldsymbol{x}}_1^{(n)}),\end{aligned}\tag{A.24}$$

令$\mathcal{S},\mathcal{S}'$交换 $\boldsymbol{x}_1^{(n)}$ 与$\tilde{\boldsymbol{x}}_1^{(n)}$，分别得到$\tilde{\mathcal{S}}_{\sigma},\tilde{\mathcal{S}'}_{\sigma}$，则此时式 (A.22)成立。

其他情况 其余5种情况与情况3类似，可证明$\mathcal{S},\mathcal{S}'$交换$\sigma_i=-1$位置对应的样例则可得到$\tilde{\mathcal{S}}_{\sigma},\tilde{\mathcal{S}'}_{\sigma}$，细节此处从略。

综上讨论，基条件得证，下面证明递推关系成立。

步骤2：考察递推关系，给定$1<n_1<n_p$，$1<n_2<n_u$，$1<n_3<n_n$、样本

$$\begin{aligned}\mathcal{S}^0&=\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_1}\cup\{\boldsymbol{x}_j^{(u)}\}_{j=1}^{n_2}\cup\{\boldsymbol{x}_k^{(n)}\}_{k=1}^{n_3}\\\mathcal{S}^{0\prime}&=\{\tilde{\boldsymbol{x}}_i^{(p)}\}_{i=1}^{n_1}\cup\{\tilde{\boldsymbol{x}}_j^{(u)}\}_{j=1}^{n_2}\cup\{\tilde{\boldsymbol{x}}_k^{(n)}\}_{k=1}^{n_3}\end{aligned}$$

以及Rademacher随机变量序列：$\sigma^0 = \{\sigma_i^{(p)}\}_{i=1}^{n_1} \cup \{\sigma_j^{(u)}\}_{j=1}^{n_2} \cup \{\sigma_k^{(n)}\}_{k=1}^{n_3}$，假设存在$\tilde{\mathcal{S}}_{\sigma^0}^0, \tilde{\mathcal{S}}_{\sigma^0}^{0\prime}$使式 (A.22) 对$\mathcal{S}^0, \mathcal{S}^{0\prime}, \sigma^0$成立。下面证明，任给新样本$\boldsymbol{x}_b^{(p)}, \tilde{\boldsymbol{x}}_b^{(p)}, \sigma_b^{(p)}$，或$\boldsymbol{x}_b^{(u)}, \tilde{\boldsymbol{x}}_b^{(u)}, \sigma_b^{(u)}$或$\boldsymbol{x}_b^{(n)}, \tilde{\boldsymbol{x}}_b^{(n)}, \sigma_b^{(n)}$，式 (A.22) 仍然成立。

显然仅$Q_\sigma^{p,n,b,k}, k = 1, 2, \cdots, n_n$, $Q_\sigma^{p,u,b,j}, j = 1, 2, \cdots, n_u$计算与新样本有关。首先考虑任意给定的$Q_\sigma^{p,n,b,k}$。先考虑$\sigma_b^{(p)} = 1$情况：

情况2-1 $\sigma_k^{(n)} = 1$，此时有：

$$Q_\sigma^{p,n,b,k} = \ell(f, \tilde{\boldsymbol{x}}_b^{(p)}, \tilde{\boldsymbol{x}}_k^{(n)}) - \ell(f, \boldsymbol{x}_b^{(p)}, \boldsymbol{x}_k^{(n)}).$$

情况2-2 $\sigma_k^{(n)} = -1$, 有:

$$Q_\sigma^{p,n,b,k} = \ell(f^{(i)}, \tilde{\boldsymbol{x}}_b^{(p)}, \boldsymbol{x}_k^{(n)}) - \ell(f, \boldsymbol{x}_b^{(p)}, \tilde{\boldsymbol{x}}_k^{(n)}).$$

同理，可以证明若$\sigma_b^{(p)} = -1$，则有：

情况3-1 $\sigma_k^{(n)} = 1$，此时有：

$$Q_\sigma^{p,n,b,k} = \ell(f, \boldsymbol{x}_b^{(p)}, \tilde{\boldsymbol{x}}_k^{(n)}) - \ell(f, \tilde{\boldsymbol{x}}_b^{(p)}, \boldsymbol{x}_k^{(n)}).$$

情况3-2 $\sigma_k^{(n)} = -1$，有:

$$Q_\sigma^{p,n,b,k} = \ell(f^{(i)}, \boldsymbol{x}_b^{(p)}, \boldsymbol{x}_k^{(n)}) - \ell(f, \tilde{\boldsymbol{x}}_b^{(p)}, \tilde{\boldsymbol{x}}_k^{(n)}).$$

对于$Q_\sigma^{p,u,b,j}$，若$\sigma_b^{(p)} = 1$，有：

情况4-1 $\sigma_j^{(u)} = 1$，此时有：$\sigma_m^{(j)} = 1$，有：

$$Q_\sigma^{p,n,b,k} = \ell(f, \tilde{\boldsymbol{x}}_b^{(p)}, \tilde{\boldsymbol{x}}_j^{(u)}) - \ell(f, \boldsymbol{x}_b^{(p)}, \boldsymbol{x}_j^{(u)}).$$

情况4-2 $\sigma_j^{(u)} = -1$，有:

$$Q_\sigma^{p,n,b,k} = \ell(f^{(i)}, \tilde{\boldsymbol{x}}_b^{(p)}, \boldsymbol{x}_j^{(u)}) - \ell(f, \boldsymbol{x}_b^{(p)}, \tilde{\boldsymbol{x}}_j^{(u)}).$$

若$\sigma_b^{(p)} = -1$，有：

情况5-1 $\sigma_j^{(u)} = 1$，此时有：

$$Q_\sigma^{p,n,b,k} = \ell(f, \boldsymbol{x}_b^{(p)}, \tilde{\boldsymbol{x}}_j^{(u)}) - \ell(f, \tilde{\boldsymbol{x}}_b^{(p)}, \boldsymbol{x}_j^{(u)}).$$

情况5-2 $\sigma_j^{(u)} = -1$，有:

$$Q_\sigma^{p,n,b,k} = \ell(f^{(i)}, \boldsymbol{x}_b^{(p)}, \boldsymbol{x}_j^{(u)}) - \ell(f, \tilde{\boldsymbol{x}}_b^{(p)}, \tilde{\boldsymbol{x}}_j^{(u)}).$$

综上讨论，若$\sigma_b^{(p)}=1$，令

$$\tilde{\mathcal{S}}_\sigma=\tilde{\mathcal{S}}^0_{\sigma^0}\cup\{\tilde{\boldsymbol{x}}_b^{(p)}\},\ \tilde{\mathcal{S}}'_\sigma=\tilde{\mathcal{S}}^{0'}_{\sigma^0}\cup\{\boldsymbol{x}_b^{(p)}\},$$

若$\sigma_b^{(p)}=-1$，令：

$$\tilde{\mathcal{S}}_\sigma=\tilde{\mathcal{S}}^0_{\sigma^0}\cup\{\boldsymbol{x}_b^{(p)}\},\ \tilde{\mathcal{S}}'_\sigma=\tilde{\mathcal{S}}^{0'}_{\sigma^0}\cup\{\tilde{\boldsymbol{x}}_b^{(p)}\},$$

可保证式 (A.22)仍然成立。

对于新样本为$\boldsymbol{x}_b^{(u)},\tilde{\boldsymbol{x}}_b^{(u)},\sigma_b^{(u)}$ 或$\boldsymbol{x}_b^{(n)},\tilde{\boldsymbol{x}}_b^{(n)},\sigma_b^{(n)}$的情况，仅需将上述$\tilde{\mathcal{S}}_\sigma$ $\tilde{\mathcal{S}}'_\sigma$构造方式中的$\sigma_b^{(p)}$替换为$\sigma_b^{(u)}$或$\sigma_b^{(n)}$，将$\tilde{\boldsymbol{x}}_b^{(p)},\boldsymbol{x}_b^{(p)}$替换为$\tilde{\boldsymbol{x}}_b^{(u)},\boldsymbol{x}_b^{(u)}$或$\tilde{\boldsymbol{x}}_b^{(n)},\boldsymbol{x}_b^{(n)}$即可保证式(A.22)成立，由于证明过程与上述内容类似，此处从略。

综上讨论，式 (A.22) 得证。基于式 (A.22) ，可进一步得出：

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\gamma}{n_pn_n}\cdot Q_\sigma^{p,n,i,k}+\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{1-\gamma}{2n_pn_u}\cdot Q_\sigma^{p,u,i,j}\right.\\
&\qquad\left.+\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{1-\gamma}{2n_un_n}\cdot Q_\sigma^{u,n,j,k}\right]\\
&=\mathop{\mathbb{E}}_{\sigma}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f)-\hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f)\right)\right]\\
&=\frac{1}{2^N}\cdot\sum_{\sigma}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f)-\hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f)\right)\right]\\
&=\frac{1}{2^N}\cdot\sum_{\sigma}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}^{PNU}_{\mathcal{S}'}(f)-\hat{R}^{PNU}_{\mathcal{S}}(f)\right)\right]\\
&=\frac{2^N}{2^N}\cdot\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}^{PNU}_{\mathcal{S}'}(f)-\hat{R}^{PNU}_{\mathcal{S}}(f)\right)\right]\\
&=\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}^{PNU}_{\mathcal{S}'}(f)-\hat{R}^{PNU}_{\mathcal{S}}(f)\right)\right].
\end{aligned}\tag{A.25}$$

由于Rademacher随机变量的符号不影响其期望及样例水平的独性，有

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}^{PNU}_{\mathcal{S}'}(f)-\hat{R}^{PNU}_{\mathcal{S}}(f)\right)\right]\le 4\mathfrak{R}_{PNU}(\ell\circ\mathcal{H}),$$

由此引理得证。 □

A.4.6　引理 A.8 证明

证明. 有$\hat{\mathfrak{R}}_{PNU}(\ell_\rho \circ co\,(\mathcal{H}_{DS}))$定义，有：

$$\hat{\mathfrak{R}}_{PNU}(\ell \circ co\,(\mathcal{H}_{DS})) \leq \underbrace{\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in co(\mathcal{H}_{DS})}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n} Q_{i,k}\right]}_{(I)}$$
$$+\underbrace{\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in co(\mathcal{H}_{DS})}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u} Q_{i,j}\right]}_{(II)}+\underbrace{\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in co(\mathcal{H}_{DS})}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n} Q_{j,k}\right]}_{(III)}$$

由引理A.4、引理A.5、ℓ_ρ函数的$\frac{1}{\rho}$–Lipschiptz连续性及上确界的次可加性，有以下结论：

$$(I) \leq \frac{\gamma}{2n_pn_n\rho}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_i^{(p)}}{2}\left(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_k^{(n)})\right)\right]$$
$$+\frac{\gamma}{n_pn_n\rho}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}\left(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_k^{(n)})\right)\right]$$

$$(II) \leq \frac{1-\gamma}{2n_pn_u\rho}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_i^{(p)}}{2}\left(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_j^{(u)})\right)\right]$$
$$+\frac{1-\gamma}{2n_pn_u\rho}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}_{DS}}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_j^{(u)}}{2}\left(f(\boldsymbol{x}_i^{(p)})-f(\boldsymbol{x}_j^{(u)})\right)\right]$$

$$(III) \leq \frac{1-\gamma}{2n_un_n\rho}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}_{DS}}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_j^{(u)}}{2}\left(f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_k^{(n)})\right)\right]$$
$$+\frac{1-\gamma}{2n_pn_n\rho}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}_{DS}}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}\left(f(\boldsymbol{x}_j^{(u)})-f(\boldsymbol{x}_k^{(n)})\right)\right]$$

定义

$$(IV) = \mathop{\mathbb{E}}_{\sigma}\left[\max_{\substack{e\in[d]\\ \theta\in\boldsymbol{T}_e}}\frac{\gamma}{n_pn_n}\left[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_i^{(p)}}{2}\left(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right]\right.$$
$$+\max_{\substack{e\in[d]\\ \theta\in\boldsymbol{T}_e}}\left[\frac{1-\gamma}{2n_pn_u}\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_j^{(u)}}{2}\left(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right]$$
$$\left.+\max_{\substack{e\in[d]\\ \theta\in\boldsymbol{T}_e}}\left[\frac{1-\gamma}{2n_un_n}\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}\left(h_\theta^e(\boldsymbol{x}_j^{(u)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right]\right]$$

$$
\begin{aligned}
(V) = \mathop{\mathbb{E}}_{\sigma}\Bigg[& \max_{\substack{e\in[d]\\ \theta\in\boldsymbol{T}_e}} \frac{\gamma}{n_p n_n}\cdot\left[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}\left(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right] \\
& + \max_{\substack{e\in[d]\\ \theta\in\boldsymbol{T}_e}} \left[\frac{1-\gamma}{2n_p n_u}\cdot\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_i^{(p)}}{2}\left(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right] \\
& + \max_{\substack{e\in[d]\\ \theta\in\boldsymbol{T}_e}} \left[\frac{1-\gamma}{2n_u n_n}\cdot\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_j^{(u)}}{2}\left(h_\theta^e(\boldsymbol{x}_j^{(u)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right]\Bigg]
\end{aligned}
$$

有：

$$
(I)+(II)+(III)\le(IV)+(IV)
$$

固定训练样本，仅将随机数σ视作随机变量，(IV)、(V)均满足引理 A.6 条件，下面由该引理分别给出其上界。首先关注(IV)，可证明

$$
\frac{\gamma}{n_p n_n}\cdot\left[\sum_{i=1}^{n_p}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}\left(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right]
$$

相对于随机变量序列$\{\sigma_i^{(p)}\}$满足有限差分性质，且对应系数为

$$
c_1=c_2=\cdots=c_{np}=\frac{\gamma}{n_p}
$$

同理可知 $\frac{1-\gamma}{2n_p n_u}\cdot\left[\sum_{i=1}^{n_p}\sum_{j=1}^{n_u}\frac{\sigma_j^{(u)}}{2}\left(h_\theta^e(\boldsymbol{x}_i^{(p)})-h_\theta^e(\boldsymbol{x}_j^{(u)})\right)\right]$满足有限差分性质，且对应系数为

$$
c_1=c_2=\cdots=c_{n_u}=\frac{1-\gamma}{2n_u}
$$

$\frac{1-\gamma}{2n_u n_n}\cdot\left[\sum_{j=1}^{n_u}\sum_{k=1}^{n_n}\frac{\sigma_k^{(n)}}{2}\left(h_\theta^e(\boldsymbol{x}_j^{(u)})-h_\theta^e(\boldsymbol{x}_k^{(n)})\right)\right]$满足有限差分性质，且对应系数为

$$
c_1=c_2=\cdots=c_{n_n}=\frac{1-\gamma}{2n_n}
$$

综合引理 B.1 及引理 A.6，令

$$
v_1=\sum_{i=1}^{n_p}\frac{\gamma^2}{n_p^2},\quad v_2=\sum_{j=1}^{n_u}\frac{(1-\gamma)^2}{4n_u^2},\quad v_3=\sum_{k=1}^{n_n}\frac{(1-\gamma)^2}{4n_n^2}
$$

由于假设集$\{h_\theta^e:e\in[d],\theta\in\boldsymbol{T}_e\}$包含$d\cdot K$个函数，有：

$$
\begin{aligned}
(IV) &\le \left(2(\log d+\log K)\cdot(v_1+v_2+v_3)\right)^{1/2} \\
&\le \left(2(\log d+\log K)\cdot\rho_\gamma(\boldsymbol{Y})\right)^{1/2}
\end{aligned}
$$

同理可证

$$
(V)\le\left(2(\log d+\log K)\cdot\rho_\gamma(\boldsymbol{Y})\right)^{1/2}
$$

综合(IV)及(V)上界，引理得证。 □

附录 B　第3章中的证明

B.1　AUC^{ova}和AUC^{ovo}的性质对比

本节对比AUC^{ova}和AUC^{ovo}，证明AUC^{ova}无法处理类别对之间的不平衡。

再次声明定理 *3.1.* 给定标签分布$\mathbb{P}[y=i]=p_i>0$以及多类得分函数f，以下性质成立:

(a)

$$\mathrm{AUC}^{\mathrm{ova}}(f)=\frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{j\neq i}\left(\frac{p_j}{1-p_i}\right)\cdot\mathrm{AUC}_{i|j}(f^{(i)}) \tag{B.1}$$

(b)

$$\mathrm{AUC}^{\mathrm{ova}}(f)=\mathrm{AUC}^{\mathrm{ovo}}(f),\ \text{当}\ \ p_i=\frac{1}{N_C},$$
$$i=1,2,\cdots,N_C$$

(c) 当且仅当$\mathrm{AUC}^{\mathrm{ovo}}(f)=1$时$\mathrm{AUC}^{\mathrm{ova}}(f)=1$。

证明.

证明(a). :对任意事件$\mathcal{C}$：

$$\mathbb{P}\left[\mathcal{C},\mathcal{E}^{(i)}\right]=\sum_{j\neq i}\mathbb{P}\left[\mathcal{C}|\mathcal{E}^{(ij)}\right]\mathbb{P}\left[\mathcal{E}^{(ij)}\right],$$

存在：

$$\begin{aligned}
\mathrm{AUC}_{i|\neg i}&=\mathbb{P}\left[(y_1^{(i)}-y_2^{(i)})\cdot(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2))>0|\mathcal{E}^{(i)}\right]\\
&\quad+\frac{1}{2}\mathbb{P}\left[f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)=0|\mathcal{E}^{(i)}\right]\\
&=\sum_{j\neq i}\frac{\mathbb{P}\left[\mathcal{E}^{(ij)}\right]}{\mathbb{P}\left[\mathcal{E}^{(i)}\right]}\cdot\left(\mathbb{P}\left[(y_1^{(i)}-y_2^{(i)})\cdot(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2))>0|\mathcal{E}^{(ij)}\right]\right.\\
&\quad\left.+\frac{1}{2}\mathbb{P}\left[f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)=0|\mathcal{E}^{(ij)}\right]\right)\\
&=\sum_{j\neq i}\frac{\mathbb{P}\left[\mathcal{E}^{(ij)}\right]}{\mathbb{P}\left[\mathcal{E}^{(i)}\right]}\mathrm{AUC}_{i|j}\\
&=\sum_{j\neq i}\frac{p_j}{1-p_i}\mathrm{AUC}_{i|j}
\end{aligned}$$

因此：

$$\begin{aligned}\mathrm{AUC}^{\mathrm{ovo}} &= \frac{1}{N_C}\sum_{i=1}^{N_C}\mathrm{AUC}_{i|\neg i}\\ &= \frac{1}{N_C}\sum_{i=1}^{N_C}\sum_{j\neq i}\frac{p_j}{1-p_i}\mathrm{AUC}_{i|j}.\end{aligned}$$

证明(b). 由(a)中的结论可证：当$p_i = \frac{1}{N_C}$, $\forall i = 1, 2, \cdots, N_C$时，$\frac{p_j}{1-p_i} = \frac{p_i}{1-p_j} = \frac{1}{N_C-1}$。

证明(c). 因为（1）$\mathrm{AUC}^{\mathrm{ovo}}$ 和 $\mathrm{AUC}^{\mathrm{ova}}$ 为 $\mathrm{AUC}_{i|j}$的凸组合；（2）$\mathrm{AUC}_{i|j} \in [0, 1]$，所以

$$\mathrm{AUC}^{\mathrm{ovo}} = 1 \leftrightarrow \mathrm{AUC}_{i|j} = 1,\ \forall i \neq j \leftrightarrow \mathrm{AUC}^{\mathrm{ova}} = 1.$$

□

B.2 一致性分析

B.2.1 贝叶斯最优评分函数

本节主要推导MAUC$^{\downarrow}$准则下的贝叶斯最优函数。

再次声明定理 *3.2.* 给定$\eta_i(\cdot) = \mathbb{P}[y = i|x]$, $p_i = \mathbb{P}[y = i]$，可得:

(a) 若

$$\Delta(f^{(i)}) \cdot \Delta(\pi) > 0,\ \forall \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \neq \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1), \tag{B.2}$$

其中：

$$\begin{aligned}\Delta(f^{(i)}) &= f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2)\\ \Delta(\pi) &= \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) - \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)\\ \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \sum_{j\neq i}\frac{\eta_i(\boldsymbol{x}_1)\eta_j(\boldsymbol{x}_2)}{2p_ip_j},\\ \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1) &= \sum_{j\neq i}\frac{\eta_j(\boldsymbol{x}_1)\eta_i(\boldsymbol{x}_2)}{2p_ip_j},\end{aligned}$$

则$f = \{f^{(i)}\}_{i=1,2,\cdots,N_C}$是MAUC$^{\downarrow}$准则下的贝叶斯最优评分函数。

(b) 令$\sigma(\cdot)$表示sigmoid函数，$s_i(\boldsymbol{x}) = \eta_i(\boldsymbol{x})/p_i$，$s_{\setminus i}(\boldsymbol{x}) = \sum_{j\neq i} s_j(\boldsymbol{x})$，则贝叶斯最优评分函数可表示为:

$$f^{\star(i)}(\boldsymbol{x}) = \begin{cases}\sigma\left(\dfrac{s_i(\boldsymbol{x})}{s_{\setminus i}(\boldsymbol{x})}\right), & 0 \leq \eta_i(\boldsymbol{x}) < 1\\ 1, & \eta_i(\boldsymbol{x}) = 1.\end{cases} \tag{B.3}$$

证明.

证明(a)

证明所需符号见表.B.1。

表 B.1 符号与描述

Table B.1 Notations and descriptions.

符号	描述
$\mathcal{Y}_{ij}$	事件$[y_1 = i, y_2 = j]$
$\mathcal{E}_{i,j}$	事件 $\mathcal{Y}_{ij}$ or $\mathcal{Y}_{ji}$
$y_1^{(i)}$	若 $y_1 = i$ 则 $= 1$，否则 $y_1^{(i)} = 0$
$y_2^{(i)}$	若 $y_2 = i$ 则 $= 1$，否则 $y_2^{(i)} = 0$
$\boldsymbol{x}_{1,2}$	$(\boldsymbol{x}_1, \boldsymbol{x}_2)$
$y_{1,2}$	(y_1, y_2)
$\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2})$	事件 $(f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2)) \cdot (y_1^{(i)} - y_2^{(i)}) < 0$
$\mathcal{G}_i\left((1,0), \boldsymbol{x}_{1,2}\right)$	事件 $(f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2)) \cdot (1 - 0) < 0$
$\mathcal{G}_i\left((0,1), \boldsymbol{x}_{1,2}\right)$	事件 $(f^{(i)}\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2)) \cdot (0 - 1) < 0$
$\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2})$	事件 $f^{(i)}(\boldsymbol{x}_1) = f^{(i)}(\boldsymbol{x}_2)$
$\eta_i(\boldsymbol{x})$	$\mathbb{P}(y = i\|\boldsymbol{x})$
$\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2)$	$\sum_{j \neq i}(\frac{\eta_i(\boldsymbol{x}_1)\eta_j(\boldsymbol{x}_2)}{2p_i p_j})$
$\pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)$	$\sum_{j \neq i}(\frac{\eta_j(\boldsymbol{x}_1)\eta_i(\boldsymbol{x}_2)}{2p_i p_j})$

首先，消除对$\mathcal{E}_{i,j}$的条件依赖。具体地：

$$\begin{aligned} N_C \cdot (N_C - 1) \cdot \mathsf{MAUC}^{\downarrow} &= \sum_{i=1}^{N_C} \sum_{j \neq i} \left(\mathbb{P}\left[\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2}) | \mathcal{E}_{i,j}\right] + \frac{1}{2} \cdot \mathbb{P}\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2}) | \mathcal{E}_{i,j}\right] \right) \\ &= \sum_{i=1}^{N_C} \sum_{j \neq i} \frac{\mathbb{P}\left[\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2}), \mathcal{E}_{i,j}\right]}{\mathbb{P}\left[\mathcal{E}_{i,j}\right]} + \frac{1}{2} \frac{\mathbb{P}\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2}), \mathcal{E}_{i,j}\right]}{\mathbb{P}\left[\mathcal{E}_{i,j}\right]} \\ &= \sum_{i=1}^{N_C} \sum_{j \neq i} \frac{\mathbb{P}\left[\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2}), \mathcal{E}_{i,j}\right]}{2p_i p_j} + \frac{1}{2} \frac{\mathbb{P}\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2}), \mathcal{E}_{i,j}\right]}{2p_i p_j} \end{aligned} \tag{B.4}$$

基于此，展开联合分布：

$$\begin{aligned} &\frac{\mathbb{P}\left[\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2}), \mathcal{E}_{i,j}\right]}{2p_i p_j} \\ &= \frac{\mathbb{E}_{\boldsymbol{x}_{1,2}, y_{1,2}}\left[\boldsymbol{I}[\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2})] \cdot \boldsymbol{I}[\mathcal{E}_{i,j}]\right]}{2p_i p_j} \\ &= \frac{\mathbb{E}_{\boldsymbol{x}_{1,2}}\left[\mathbb{E}_{y_{1,2}|\boldsymbol{x}_{1,2}}\left[\boldsymbol{I}[\mathcal{G}_i(y_{1,2}, \boldsymbol{x}_{1,2})] \cdot \boldsymbol{I}[\mathcal{E}_{i,j}]\right]\right]}{2p_i p_j} \\ &= \frac{\mathbb{E}_{\boldsymbol{x}_{1,2}}\left[\eta_i(\boldsymbol{x}_1) \cdot \eta_j(\boldsymbol{x}_2) \cdot \boldsymbol{I}[\mathcal{G}_i\left((1,0), \boldsymbol{x}_{1,2}\right)] + \eta_j(\boldsymbol{x}_1) \cdot \eta_i(\boldsymbol{x}_2) \cdot \boldsymbol{I}[\mathcal{G}_i\left((0,1), \boldsymbol{x}_{1,2}\right)]\right]}{2p_i p_j}. \end{aligned} \tag{B.5}$$

类似地：

$$\frac{\mathbb{P}\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2}), \mathcal{E}_{i,j})\right]}{2p_i p_j} = \frac{\mathbb{E}_{\boldsymbol{x}_{1,2}}\left[\eta_i(\boldsymbol{x}_1) \cdot \eta_j(\boldsymbol{x}_2) \cdot \boldsymbol{I}\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2})\right] + \eta_j(\boldsymbol{x}_1) \cdot \eta_i(\boldsymbol{x}_2) \cdot \boldsymbol{I}\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2})\right]\right]}{2p_i p_j} \tag{B.6}$$

综上，组合式 (B.4)-式 (B.6)可得：

$$\begin{aligned} N_C \cdot (N_C - 1) \cdot \mathsf{MAUC}^{\downarrow} = \sum_{i=1}^{N_C} \mathop{\mathbb{E}}_{x_{1,2}} \Big[& \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot \boldsymbol{I}\left[\mathcal{G}_i\left((1,0), \boldsymbol{x}_{1,2}\right)\right] \\ & + \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)) \cdot \boldsymbol{I}\left[\mathcal{G}_i\left((0,1), \boldsymbol{x}_{1,2}\right)\right] \\ & + \frac{1}{2} \cdot \left(\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) + \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)\right) \cdot I\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2})\right] \Big]. \end{aligned} \tag{B.7}$$

固定$f^{(i)}$和$\boldsymbol{x}_{1,2}$，只需最小化$L^{(i)}$即可得到贝叶斯最优解：

$$\begin{aligned} L^{(i)} =& \pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\cdot \boldsymbol{I}\left[\mathcal{G}_i\left((1,0),\boldsymbol{x}_{1,2}\right)\right] \\ &+\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1))\cdot \boldsymbol{I}\left[\mathcal{G}_i\left((0,1),\boldsymbol{x}_{1,2}\right)\right]+\frac{1}{2}\cdot\left(\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)+\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)\right)\cdot I\left[\mathcal{G}_{0,i}(\boldsymbol{x}_{1,2})\right] \end{aligned} \tag{B.8}$$

显然：

$$L^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)=\begin{cases} \pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2), & f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)<0, \\ \dfrac{\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)+\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)}{2}, & f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)=0, \\ \pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1), & f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)>0. \end{cases} \tag{B.9}$$

对任意$\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}$，最小化$L(\boldsymbol{x}_1,\boldsymbol{x}_2)$即可得到贝叶斯最优评分函数：

若$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)=\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$，则$L(\boldsymbol{x}_1,\boldsymbol{x}_2)$是与$f^{(i)}$无关的常数。

若$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)<\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$，则必须在满足$f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)<0$的条件下最小化$L(\boldsymbol{x}_1,\boldsymbol{x}_2)$。

若$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)>\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$，则必须在满足$f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)>0$的条件下最小化$L(\boldsymbol{x}_1,\boldsymbol{x}_2)$。

综上，通过选取满足（1）$\forall\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}$；（2）$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\neq\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$；（3）$\left(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)\right)\cdot\left(\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)-\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)\right)>0$ 的$f^{(i)}$，即可得到$R^{(i)}(f^{(i)})$最小值为 $\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\min\{\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2),\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)\}$.

证明(b).

固定类别i，若$\eta_i(\boldsymbol{x}_k)\neq 1,\ k=1,2$，则通过对定理.3.2中的公式.(B.2) 除以$s_{\setminus i}(\boldsymbol{x}_1)\cdot s_{\setminus i}(\boldsymbol{x}_2)$可知，可选择$\sigma(s_i(\boldsymbol{x})/s_{\setminus i}(\boldsymbol{x}))$作为$f^{\star(i)}(\boldsymbol{x})$。

若$\eta_i(\boldsymbol{x}_1)=1,\eta_i(\boldsymbol{x}_2)<1$，（注意：$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\neq\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$) 则 $s_{\setminus i}(\boldsymbol{x}_1)=0$。为得到$f^{\star(i)}(\boldsymbol{x}_1)>f^{\star(i)}(\boldsymbol{x}_2)$，设 $f^{\star(i)}(\boldsymbol{x}_1)=1$，$f^{\star(i)}(\boldsymbol{x}_2)=\sigma\left(s_i(\boldsymbol{x}_2)/s_{\setminus i}(\boldsymbol{x}_2)\right)$。

在后一个例子中，可得到$\eta_i(\boldsymbol{x}_1)<1,\eta_i(\boldsymbol{x}_2)=1$。由此可知最优解能够满足$f(\boldsymbol{x}_1)=\sigma\left(s_i(\boldsymbol{x}_1)/s_{\setminus i}(\boldsymbol{x}_1)\right),f(\boldsymbol{x}_2)=1$.

由于$\boldsymbol{x}_1$，$\boldsymbol{x}_2$可任意选择，当$\eta_i(\boldsymbol{x})<1$时，贝叶斯评分函数可以被设为: $f^{\star(i)}(\boldsymbol{x})=\sigma(s_i(\boldsymbol{x})/s_{\setminus i}(\boldsymbol{x}))(<1)$；否则$f^{\star(i)}(\boldsymbol{x})=1$。 □

B.2.2 替代损失的一致性

基于定理.3.2，本节对一些主流替代损失给出理论分析。

再次声明定理3.3. 若体大损失函数ℓ满足（1）可微；（2）为凸函数；（3）在$[-1,1]$内非递增；（4）$\ell'(0)<0$，则该替代损失和MAUC$^{\downarrow}$一致。

证明. 首先回顾替代风险的定义：

$$R_\ell(f)=\sum_i \frac{R_\ell^{(i)}(f^{(i)})}{N_C(N_C-1)}$$
$$R_\ell^{(i)}(f^{(i)})=\sum_{j\neq i}\mathop{\mathbb{E}}_{z_1,z_2}\left[\ell(\Delta(y_{1,2}^{(i)})\Delta f^{(i)})|\mathcal{E}^{(ij)}\right],$$

其中 $R_\ell^{(i)}(f^{(i)})$ 是 $f^{(i)}$ 对应的风险。

记：

$$f^\star=\mathop{\operatorname{argmin}}_{f\in\mathcal{F}_\sigma} R_\ell(f)$$

且 $f^\star=(f^{\star(1)},f^{\star(2)},\cdots,f^{\star(N_C)})$。

此外，以下等式显然成立：

$$\inf_{f\in\mathcal{F}_\sigma} R_\ell(f)=\frac{1}{N_C\cdot(N_C-1)}\cdot\sum_i \inf_{f^{(i)}\in\mathcal{F}_\sigma} R_\ell^{(i)}(f^{(i)}). \tag{B.10}$$

基于此，每个二分类评分函数 $f^{\star(i)}$ 都可以通过其对应的子问题单独求解，即

$$f^{\star(i)}=\mathop{\operatorname{argmin}}_{f^{(i)}\in\mathcal{F}_\sigma} R_\ell^{(i)}(f^{(i)})$$

和定理.3.2的推导类似，将$R_\ell^{(i)}$整理为：

$$R_\ell^{(i)}(f^{(i)})=\frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}}\Big[\eta_+^{(i)}(\boldsymbol{x})\cdot\eta_-^{(i)}(\boldsymbol{x}')\cdot\ell(f^{(i)}(\boldsymbol{x})-f^{(i)}(x'))$$
$$+\eta_+^{(i)}(\boldsymbol{x}')\cdot\eta_-^{(i)}(\boldsymbol{x})\cdot\ell(f^{(i)}(\boldsymbol{x}')-f^{(i)}(\boldsymbol{x}))\Big]d\mathbb{P}(\boldsymbol{x})d\mathbb{P}(\boldsymbol{x}')$$

其中

$$\eta_+^{(i)}(\boldsymbol{x})=\frac{\eta_i(\boldsymbol{x})}{p_i},\ \eta_-^{(i)}(\boldsymbol{x})=\sum_{j\neq i}\frac{\eta_j(\boldsymbol{x})}{p_j}。$$

记$\text{Bayes}_{\sigma}^{(i)}$为$f^{(i)}$的集合，其中 f 为贝叶斯最优函数, 即

$$\text{Bayes}_{\sigma}^{(i)} =\{f^{(i)} : \left(f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2)\right) \cdot (\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) - \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)) > 0,$$
$$\forall(\boldsymbol{x}_1, \boldsymbol{x}_2)\ s.t.\ \pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) \neq \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)\}$$

基于此，可通过以下步骤证明此定理。

声明 1. $f^{\star(i)} \in \text{Bayes}_{\sigma}^{(i)}$

可通过反证法证明以上声明。首先假设该声明在以下场景中不成立：$\exists \boldsymbol{x}_1, \boldsymbol{x}_2$ 满足 $f^{\star(i)}(\boldsymbol{x}_1) \leq f^{\star(i)}(\boldsymbol{x}_2)$ 但 $\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) > \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)$。

情况 1: $\eta_{-}^{(i)}(\boldsymbol{x}_1) > 0, \eta_{-}^{(i)}(\boldsymbol{x}_2) > 0$

定义 $\delta_h(\gamma) = R_{\ell}^{(i)}(f^{\star(i)} + \gamma h)$，必满足：$\delta_h'(0) = 0, \forall h$。

给定任意一对样本 $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}$（其中 $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$），需证明：对所有满足充分条件的$\ell$，通过反证法即可由$f^{\star(i)}(\boldsymbol{x}_1) > f^{\star(i)}(\boldsymbol{x}_2)$ 推出$\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) > \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)$。

选择

$$h_1(x) = \frac{\boldsymbol{I}[x = x_1]}{\eta_{-}(\boldsymbol{x})},$$

由$\delta_{h_1}'(0) = 0$可得

$$\begin{aligned}\int_{X \setminus \boldsymbol{x}_1} \Bigg[& \frac{\eta_{+}^{(i)}(\boldsymbol{x}_1)}{\eta_{-}^{(i)}(\boldsymbol{x}_1)} \cdot \eta_{-}^{(i)}(\boldsymbol{x}) \cdot \ell' \left(f^{\star(i)}(\boldsymbol{x}_1) - f^{\star(i)}(\boldsymbol{x})\right) \\ & - \eta_{+}^{(i)}(\boldsymbol{x}) \cdot \ell' \left(f^{\star(i)}(\boldsymbol{x}) - f^{\star(i)}(\boldsymbol{x}_1)\right) \Bigg] d\mathbb{P}(\boldsymbol{x}) \\ = & \ 0\end{aligned} \tag{B.11}$$

选择

$$h_2(x) = \frac{\boldsymbol{I}[x = x_2]}{\eta_{-}(\boldsymbol{x})},$$

由$\delta_{h_2}'(0) = 0$可得：

$$\begin{aligned}\int_{X \setminus \boldsymbol{x}_2} \Bigg[& \frac{\eta_{+}^{(i)}(\boldsymbol{x}_2)}{\eta_{-}^{(i)}(\boldsymbol{x}_2)} \cdot \eta_{-}^{(i)}(\boldsymbol{x}) \cdot \ell' \left(f^{\star(i)}(\boldsymbol{x}_2) - f^{\star(i)}(\boldsymbol{x})\right) \\ & - \eta_{+}^{(i)}(\boldsymbol{x}) \cdot \ell' \left(f^{\star(i)}(\boldsymbol{x}) - f^{\star(i)}(\boldsymbol{x}_2)\right) \Bigg] d\mathbb{P}(\boldsymbol{x}) \\ = & \ 0\end{aligned} \tag{B.12}$$

通过式 (B.11) - 式 (B.12)可得：

$$
\begin{aligned}
&\underbrace{\int_{X\setminus\{\boldsymbol{x}_1,\boldsymbol{x}_2\}} \eta_+^{(i)}(\boldsymbol{x})\cdot\left(\ell'\left(f^{\star(i)}(\boldsymbol{x})-f^{\star(i)}(\boldsymbol{x}_2)\right)-\ell'\left(f^{\star(i)}(\boldsymbol{x})-f^{\star(i)}(\boldsymbol{x}_1)\right)\right)d\mathbb{P}(\boldsymbol{x})}_{(\boldsymbol{a})}\\
&+\underbrace{\int_{X\setminus\{\boldsymbol{x}_1,\boldsymbol{x}_2\}} \frac{\eta_-^{(i)}(\boldsymbol{x})}{\eta_-^{(i)}(\boldsymbol{x}_2)\eta_-^{(i)}(\boldsymbol{x}_1)}\left(\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\cdot\ell'\left(f^{\star(i)}(\boldsymbol{x}_1)-f^{\star(i)}(\boldsymbol{x})\right)-\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)\cdot\ell'\left(f^{\star(i)}(\boldsymbol{x}_2)-f^{\star(i)}(\boldsymbol{x})\right)\right)d\mathbb{P}(\boldsymbol{x})}_{(\boldsymbol{b})}\\
&+\underbrace{\left(\frac{\mathbb{P}(\boldsymbol{x}_1)}{\eta_-^{(i)}(\boldsymbol{x}_1)}+\frac{\mathbb{P}(\boldsymbol{x}_2)}{\eta_-^{(i)}(\boldsymbol{x}_2)}\right)\cdot\left(\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\cdot\ell'\left(f^{\star(i)}(\boldsymbol{x}_1)-f^{\star(i)}(\boldsymbol{x})\right)-\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)\cdot\ell'\left(f^{\star(i)}(\boldsymbol{x}_2)-f^{\star(i)}(\boldsymbol{x})\right)\right)}_{(\boldsymbol{c})}\\
&=0
\end{aligned}
\tag{B.13}
$$

基于以上各式即可构建反证法中的矛盾。

首先注意:

$$\ell'\left(f^{\star(i)}(\boldsymbol{x})-f^{\star(i)}(\boldsymbol{x}_2)\right)\leq\ell'\left(f^{\star(i)}(\boldsymbol{x})-f^{\star(i)}(\boldsymbol{x}_1)\right)$$

其原因在于：由于假设 ℓ 非凸, 即, ℓ' 非递减。由此可知$(\boldsymbol{a})\leq 0$。

类似可知：

$$\ell'\left(f^{\star(i)}(\boldsymbol{x}_1)-f^{\star(i)}(\boldsymbol{x})\right)\leq\ell'\left(f^{\star(i)}(\boldsymbol{x}_2)-f^{\star(i)}(\boldsymbol{x})\right)\leq 0.$$

再由$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)>\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$，可得$(\boldsymbol{b})\leq 0$。

接下来证明 $(\boldsymbol{c})<0$：

情况 a: 若 $f^{\star(i)}(\boldsymbol{x}_1)=f^{\star(i)}(\boldsymbol{x}_2)$，则

$$(\boldsymbol{c})=(\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)-\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1))\cdot\ell'(0)<0.$$

原因在于$\ell(0)<0$。

情况 b: 若 $f^{\star(i)}(\boldsymbol{x}_1)<f^{\star(i)}(\boldsymbol{x}_2)$，则

$$\ell'\left(f^{\star(i)}(\boldsymbol{x}_1)-f^{\star(i)}(\boldsymbol{x}_2)\right)\leq\ell'(0)<0,$$

$$\ell'\left(f^{\star(i)}\left(\boldsymbol{x}_2\right)-f^{\star(i)}\left(\boldsymbol{x}_1\right)\right)\leq 0,$$

且$\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)>\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)$。由此可得$\boldsymbol{(c)}<0$

综上可知 $\boldsymbol{(a)}+\boldsymbol{(b)}+\boldsymbol{(c)}<0$。该结论和公式.式 (B.14)相矛盾。由于可选择任意$(\boldsymbol{x}_1,\boldsymbol{x}_2)$，可知

$$f^{\star(i)}\in \mathsf{Bayes}_{\sigma}^{(i)},\ \text{if}\ \eta_{-}^{(i)}(\boldsymbol{x}_1)>0,\eta_{-}^{(i)}(\boldsymbol{x}_2)>0$$

情况 2: $\eta_{-}^{(i)}(\boldsymbol{x}_1)=0,\eta_{-}^{(i)}(\boldsymbol{x}_2)>0$

由

$$\sum_i \mathbb{P}\left[\boldsymbol{x}_1,y=i\right]=\sum_i \eta_i(\boldsymbol{x}_1)\mathbb{P}(\boldsymbol{x}_1)=\mathbb{P}(\boldsymbol{x}_1)$$

可得：

$$\sum_i \eta_i(\boldsymbol{x}_1)=1.$$

该结论进一步表明$\eta_i(\boldsymbol{x}_1)$ 达到了其最小值1。由此可知$\eta_{+}^{(i)}(\boldsymbol{x}_1)>\eta_{+}^{(i)}(\boldsymbol{x}_2)$。

和情况1的证明类似，可令 $h_1(\boldsymbol{x})=\boldsymbol{I}\left[\boldsymbol{x}=\boldsymbol{x}_1\right]$，$h_2(\boldsymbol{x})=\boldsymbol{I}\left[\boldsymbol{x}=\boldsymbol{x}_2\right]$ 进而得到以下不等式。

$$\begin{aligned}
&\underbrace{\int_{X\setminus\{\boldsymbol{x}_1,\boldsymbol{x}_2\}}\eta_{-}^{(i)}(\boldsymbol{x})\cdot\left(\eta_{+}^{(i)}(\boldsymbol{x}_1)\cdot\ell'\left(f^{\star(i)}\left(\boldsymbol{x}\right)-f^{\star(i)}\left(\boldsymbol{x}_1\right)\right)-\eta_{+}^{(i)}(\boldsymbol{x}_2)\cdot\ell'\left(f^{\star(i)}\left(\boldsymbol{x}\right)-f^{\star(i)}\left(\boldsymbol{x}_2\right)\right)\right)d\mathbb{P}(\boldsymbol{x})}_{(\boldsymbol{a})}\\
&+\underbrace{\int_{X\setminus\{\boldsymbol{x}_1,\boldsymbol{x}_2\}}\eta_{+}^{(i)}(\boldsymbol{x})\cdot\eta_{-}^{(i)}(\boldsymbol{x}_2)\cdot\ell'\left(f^{\star(i)}\left(\boldsymbol{x}_2\right)-f^{\star(i)}\left(\boldsymbol{x}\right)\right)d\mathbb{P}(\boldsymbol{x})}_{(\boldsymbol{b})}\\
&+\underbrace{(\mathbb{P}(\boldsymbol{x}_1)+\mathbb{P}(\boldsymbol{x}_2))\cdot\eta_{+}^{(i)}(\boldsymbol{x}_1)\cdot\eta_{-}^{(i)}(\boldsymbol{x}_2)\cdot\ell'\left(f^{\star(i)}\left(\boldsymbol{x}_1\right)-f^{\star(i)}\left(\boldsymbol{x}_2\right)\right)}_{(\boldsymbol{c})}\\
&=0
\end{aligned}\tag{B.14}$$

由此可知 $\boldsymbol{(a)}+\boldsymbol{(b)}+\boldsymbol{(c)}<0$。该结论于最优条件矛盾。

$$f^{\star(i)}\in \mathsf{Bayes}_{\sigma}^{(i)},\ \text{if}\ \eta_{-}^{(i)}(\boldsymbol{x}_1)=0,\eta_{-}^{(i)}(\boldsymbol{x}_2)>0$$

情况 3: $\eta_{-}^{(i)}(\boldsymbol{x}_1)>0,\eta_{-}^{(i)}(\boldsymbol{x}_2)=0$. 在此情况下$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)<\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$与假设$\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)>\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)$矛盾。 **因此不可能出现。**

情况 4: $\eta_{-}^{(i)}(\boldsymbol{x}_1) = 0, \eta_{-}^{(i)}(\boldsymbol{x}_2) = 0$ 在此情况下$\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)$与假设$\pi^{(i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) > \pi^{(i)}(\boldsymbol{x}_2, \boldsymbol{x}_1)$矛盾。 **因此不可能出现。**

声明1证明完毕。

声明 2.

$$\inf_{f^{(i)} \notin \mathsf{Bayes}_\sigma} R_\ell^{(i)}(f^{(i)}) > \inf_{f^{(i)} \in \mathcal{F}_\sigma} R_\ell^{(i)}(f^{(i)}) \tag{B.15}$$

由声明2可得： $\operatorname{argmin}_{f^{(i)} \in \mathcal{F}_\sigma} R_\ell^{(i)}(f^{(i)}) = \mathsf{Bayes}_\sigma^{(i)}$.

声明 3. 对满足$f_i \in \mathcal{F}_\sigma$的任意序列$\{f_t\}_{t \in \mathbb{N}_+}$有：

$$R_\ell^{(i)}(f_t) \to R_\ell^{(i)}(f^{\star(i)}) \text{ implies } R^{(i)}(f_t) \to \inf_{f \in \mathcal{F}_\sigma} R^{(i)}(f)$$

由于 $\operatorname{argmin}_{f \in \mathcal{F}_\sigma} R^{(i)}(f) = \mathsf{Bayes}_\sigma^{(i)}$，只需证明$\lim_{t \to \infty} f_t \in \mathsf{Bayes}_\sigma^{(i)}$。定义 $\delta^{(i)}$为：

$$\delta^{(i)} = \inf_{f^{(i)} \notin \mathsf{Bayes}_\sigma^{(i)}} R_\ell^{(i)}(f^{(i)}) - \inf_{f^{(i)} \in \mathcal{F}_\sigma} R_\ell^{(i)}(f^{(i)}) > 0$$

假设$\lim f_t \notin \mathcal{F}_\sigma$，则对于足够大的$T$有：

$$R_\ell^{(i)}(f_T) - R_\ell^{(i)}(f^{\star(i)}) > \delta^{(i)}\Theta$$

该结论与$R_\ell^{(i)}(f_t) \to R_\ell^{(i)}(f^{\star(i)})$相矛盾。由此可知式 (B.15)成立。

由于以上结论对任意类别i都成立，因此式 (B.15)对所有i都成立。

声明 4. 给定一个序列$\{f\}_t$，其中 $f_t = (f_t^{(1)}, \cdots, f_t^{(N_C)})$，以下式子成立：

$$R_\ell(f_t) \to R_\ell(f) \text{ implies } R(f_t) \to \inf_{f \in \mathcal{F}_\sigma} R(f)$$

由此可推出：

$$R(f_t) = \frac{\sum_{i=1}^{N_C} R^{(i)}(f_t^{(i)})}{N_C \cdot (N_C - 1)},\ R_\ell(f_t) = \frac{\sum_{i=1}^{N_C} R_\ell^{(i)}(f_t^{(i)})}{N_C \cdot (N_C - 1)}.$$

以及声明4.

根据定义.3.1，证明完毕。 □

B.3　R_{surr}的无偏估计

本节推导R_{surr}的无偏估计：

再次声明性质 *3.1.* 定义$\hat{R}_{\ell,\mathcal{S}}(f)$为:

$$\hat{R}_{\ell,\mathcal{S}}(f) = \sum_{i=1}^{N_C} \sum_{j \neq i} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \ell_{surr}^{i,j,m,n},$$

其中$\ell_{surr}^{i,j,m,n}$为$\ell_{surr}(f^{(i)}(\boldsymbol{x}_m) - f^{(i)}(\boldsymbol{x}_n))$的简写；$\hat{R}_{\ell,\mathcal{S}}(f)$为$R_{surr}(f)$的无偏估计，满足 $R_{surr}(f) = \mathop{\mathbb{E}}_{\mathcal{S}}(\hat{R}_{\ell,\mathcal{S}}(f))$。

证明.

$$\mathop{\mathbb{E}}_{\mathcal{S}}(\hat{R}_{\ell,\mathcal{S}}(f)) = \mathop{\mathbb{E}}_{\boldsymbol{Y}}\left[\mathop{\mathbb{E}}_{\boldsymbol{X}|\boldsymbol{Y}}(\hat{R}_{\ell,\mathcal{S}}(f))\right]$$

此外：

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{X}|\boldsymbol{Y}}(\hat{R}_{\ell,\mathcal{S}}(f)) \\
&= \frac{1}{N_C(N_C-1)} \sum_{i=1}^{N_C} \sum_{j \neq i} \sum_{\boldsymbol{x}_1 \in \mathcal{N}_i} \sum_{\boldsymbol{x}_2 \in \mathcal{N}_j} \frac{1}{n_i n_j} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\ell_{surr}(f^{(i)}, \boldsymbol{x}_1, \boldsymbol{x}_2) | y_1 = i, y_2 = j\right] \\
&= \frac{1}{N_C(N_C-1)} \sum_{i=1}^{N_C} \sum_{j \neq i} \mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\ell_{surr}(f^{(i)}, \boldsymbol{x}_1, \boldsymbol{x}_2) | y_1 = i, y_2 = j\right].
\end{aligned}$$

由于 $\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\ell_{surr}(f^{(i)}, \boldsymbol{x}_1, \boldsymbol{x}_2) | y_1 = i, y_2 = j\right]$ 不依赖$\boldsymbol{Y}$的分布，因此：

$$\mathop{\mathbb{E}}_{\boldsymbol{Y}}\left[\mathop{\mathbb{E}}_{\boldsymbol{X}|\boldsymbol{Y}}(\hat{R}_{\ell,\mathcal{S}}(f))\right] = \mathop{\mathbb{E}}_{\boldsymbol{X}|\boldsymbol{Y}}(\hat{R}_{\ell,\mathcal{S}}(f)).$$

基于此，只需证明

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\ell_{surr}\left(f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2)\right) | y_1 = i, y_2 = j\right] = \mathop{\mathbb{E}}_{z_1,z_2}\left[\ell_{surr}\left(\Delta(y^{(i)})\Delta(f^{(i)})\right) | \mathcal{E}^{(ij)}\right].$$

为完成该证明，首先得到：

$$\mathop{\mathbb{E}}_{z_1,z_2}\left[\ell_{surr}\left(\Delta(y^{(i)})\Delta(f^{(i)})\right)|\mathcal{E}^{(ij)}\right]$$

$$\overset{(a)}{=}\mathop{\mathbb{E}}_{z_1,z_2}\left[\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\ell_{surr}\left(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)\right)+\pi^{(i)}(\boldsymbol{x}_2,\boldsymbol{x}_1)\ell_{surr}\left(f^{(i)}(\boldsymbol{x}_2)-f^{(i)}(\boldsymbol{x}_1)\right)\right]$$

$$=2\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\pi^{(i)}(\boldsymbol{x}_1,\boldsymbol{x}_2)\ell_{surr}\left(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)\right)\right]$$

$$=\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\frac{\eta^{(i)}(\boldsymbol{x}_1)\eta^{(j)}(\boldsymbol{x}_2)}{p_ip_j}\ell_{surr}\left(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)\right)\right]$$

$$\overset{(b)}{=}\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[\ell_{surr}\left(f^{(i)}(\boldsymbol{x}_1)-f^{(i)}(\boldsymbol{x}_2)\right)|y_1=i,y_2=j\right].$$

其中(a) 和 (b)可由引理.3.2得到。证明完毕。 □

B.4 泛化分析的准备工作

B.4.1 集中不等式（Concentration Inequalities）

B.4.1.1 有限差分性质（Bounded Difference Property）

定义 B.1 (有限差分性质). 给定独立随机变量$X_1,\cdots,X_n$且$X_i\in\mathbb{X}$，若存在非负常数$c_1,c_2,\cdots,c_n$满足：

$$\sup_{x_1,x_2,\cdots,x_n,x_i'}\left|f(x_1,\cdots,x_n)-f(x_1,\cdots,x_{i-1},x_i',\cdots,x_n)\right|\leq c_i,\ \forall 1\leq i\leq n. \tag{B.16}$$

则称函数$f(X_1,X_2,\cdots,X_n)$具备有限差分性质。

对所有满足有限差分性质的函数，其矩生成函数满足以下有限差分不等式（Bounded Difference Inequality）。

引理 B.1 (有限差分不等式). (Boucheron 等, 2013, 性质.6.1与定理.6.2) 假设$Z=f(X_1,\cdots,X_n)$其中各X_i独立，且满足以 $c_1,c_2,\cdots,c_n$为常数的有限差分性质。令

$$v=\frac{1}{4}\sum_{i=1}^{n}c_i^2 \tag{B.17}$$

则：

$$\log\mathbb{E}\left[\exp\left(\lambda(Z-\mathbb{E}[Z])\right)\right]\leq\frac{\lambda v^2}{2}, \tag{B.18}$$

对于所有$\lambda>0$成立。

B.4.1.2　极大值不等式

引理 B.2 (极大值不等式). (Boucheron 等, 2013, 第2.5节) 令 $Z_1, \cdots, Z_n$ 为实数随机变量，$v > 0$，且对于任意$i = 1, 2, \cdots, n$有 $\log\left(\mathbb{E}\left[\exp\left(\lambda Z_i\right)\right]\right) \leq \frac{\lambda v^2}{2}$，则：

$$\mathbb{E}\left[\max_{i=1,2,\cdots,n} Z_i\right] \leq \sqrt{2v \log n}.$$

B.4.1.3　Mcdiarmid不等式

首先回顾证明所需的集中不等式。

引理 B.3 (Mcdiarmid不等式). (McDiarmid, 1998) 令 $X_1, \cdots, X_m$ 为从 $\mathcal{X}$ 取值的独立随机变量。令 $f : \mathcal{X} \to \mathbb{R}$ 为 $X_1, \cdots, X_m$的函数且满足：

$$\sup_{\boldsymbol{x}, \boldsymbol{x}'} |f(x_1, \cdots, x_i, \cdots, x_m) - f(x_1, \cdots, x_i' \cdots, x_m)| \leq c_i,$$

其中 $\boldsymbol{x} \neq \boldsymbol{x}'$；那么对所有$\epsilon > 0$，

$$\mathbb{P}[\mathbb{E}(f) - f \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

B.4.2　Rademacher Averages的性质

要对向量-值函数求Rademacher复杂度，首先引入(Maurer, 2016; Cortes 等, 2016)中证明的向量版Talagrand收缩引理。

引理 B.4 (向量版Talagrand收缩引理). 令 $\mathcal{X}$ 为任意集合 $n \in \mathcal{N}$，$(\boldsymbol{x}_1; \cdots; \boldsymbol{x}_N) \in X_n$，令$\mathcal{F}$ 代表一类具有k各分量的函数 $f = (f^{(1)}, \cdots, f^{(k)}) : X \to \ell_2^k$ 并且令$h : \ell_2^k \to \mathbb{R}$具有Lipschitz范数$\phi$。那么，

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^N \sigma_i h(f(\boldsymbol{x}_i))\right] \leq \sqrt{2}\phi \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^N \sum_{k=1}^K \sigma_{i,k}(f^{(k)}(\boldsymbol{x}_i))\right],$$

其中 $\sigma_1 \cdots, \sigma_N$，并且 $\sigma_{1,1}, \cdots, \sigma_{1,K}, \cdots, \sigma_{N,K}$ 是两个由独立的Rademacher随机变量组成的序列。

以下引理为(Golowich 等, 2018)中引理.1在双下标Rademacher随机序列上的扩展。

引理 B.5. 令s满足1-Lipschitz且正齐次的逐元素激活函数，则对于任意类向量-值函数$\mathcal{F}$以及任意单调增的凸函数 $g:\mathbb{R}\to[0,\infty)$：

$$\begin{aligned}&\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F},||W||_F\le R_W} g\left(\left\|\sum_{i=1}^{N}\sum_{c=1}^{N_C}\sigma_{i,c}\cdot s(\boldsymbol{W}f(\boldsymbol{x}_i))\right\|\right)\right]\\ &\le 2\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} g\left(R_W\cdot\left\|\sum_{i=1}^{N}\sum_{c=1}^{N_C}\sigma_{i,c}\cdot f(\boldsymbol{x}_i)\right\|\right)\right]\end{aligned}$$

其中 $\{\sigma_{i,c}\}$ 为双下标Rademacher随机变量组成的序列。

B.4.3 Softmax的Lipschitz性质

引理 B.6 (softmax组件的Lipschitz常数). 给定 $\mathcal{X}\in\mathbb{R}^K$，函数$\mathsf{soft}_i$, $i=1,2,\cdots,K$ 为映射 $\boldsymbol{x}=(x_1,\cdots,x_k)\in\mathcal{X}\to[0,1]$，被定义为：

$$\mathsf{soft}_i(\boldsymbol{x})=\frac{\exp(x_i)}{\sum_{j=1}^{K}\exp(x_j)},$$

因此 $\mathsf{soft}_i(\cdot)$ 对于向量ℓ_2范数满足 $\frac{\sqrt{2}}{2}$-Lipschitz连续。

证明. 只需证明

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|\nabla_{\boldsymbol{x}}\mathsf{soft}_i(\boldsymbol{x})\right\|_2\le\frac{\sqrt{2}}{2}.$$

对任意 $\boldsymbol{x}\in\mathcal{X}$：

$$\frac{\partial\mathsf{soft}_i(\boldsymbol{x})}{\partial x_j}=\mathsf{soft}_i(\boldsymbol{x})\cdot\left(I[i=j]-\mathsf{soft}_j(\boldsymbol{x})\right),\ i,j=1,\cdots,K.$$

因此有：

$$\left\|\nabla_{\boldsymbol{x}}\mathsf{soft}_i(\boldsymbol{x})\right\|_2=\left(\mathsf{soft}_i^2(\boldsymbol{x})\cdot\sum_{j\neq i}\mathsf{soft}_j^2(\boldsymbol{x})+\mathsf{soft}_i(\boldsymbol{x})^2\cdot\left(1-\mathsf{soft}_j(\boldsymbol{x})\right)^2\right)^{1/2}.$$

由于 $\mathsf{soft}_i^2(\boldsymbol{x})\le\mathsf{soft}_i(\boldsymbol{x})$，$\left(1-\mathsf{soft}_j(\boldsymbol{x})\right)^2\le\left(1-\mathsf{soft}_j(\boldsymbol{x})\right)$，和 $\sum_{j\neq i}\mathsf{soft}_j^2(\boldsymbol{x})\le(1-\mathsf{soft}_i(\boldsymbol{x}))$，可得：

$$\left\|\nabla_{\boldsymbol{x}}\mathsf{soft}_i(\boldsymbol{x})\right\|_2\le\left(2\cdot\mathsf{soft}_i(\boldsymbol{x})\cdot(1-\mathsf{soft}_i(\boldsymbol{x}))\right)^{1/2}\le\frac{\sqrt{2}}{2},$$

证明完毕。

B.5 MAUC$^{\downarrow}$Rademacher复杂度及其性质

B.5.1 MAUC$^{\downarrow}$对称性

本节给出MAUC$^{\downarrow}$对称性的推导过程，该性质是证明定理.B.5.2的关键。

引理 B.7 (MAUC$^{\downarrow}$ 对称性). 以下不等式成立：

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}'}(f)\right)\right]\leq\frac{4\mathfrak{R}_{\mathrm{MAUC}^{\downarrow}}(\ell\circ\mathcal{H})}{N_C(N_C-1)},$$

其中标签$\boldsymbol{Y}$对于$\mathcal{S}$ 和 $\mathcal{S}'$固定。

证明. 记 $\ell(f,\boldsymbol{x},\boldsymbol{x}')=\ell(f^{(i)}(\boldsymbol{x}_m)-f^{(i)}(\boldsymbol{x}'_n))$ 并定义

$$\begin{aligned}T^{i,j,m,n}=&\frac{\sigma_m^{(i)}+\sigma_n^{(j)}}{2}\ell(f,\boldsymbol{x}',\boldsymbol{x}')+\frac{\sigma_m^{(i)}-\sigma_n^{(j)}}{2}\ell(f,\boldsymbol{x}',\boldsymbol{x})\\&-\frac{\sigma_m^{(i)}-\sigma_n^{(j)}}{2}\ell(f,\boldsymbol{x},\boldsymbol{x}')-\frac{\sigma_m^{(i)}+\sigma_n^{(j)}}{2}\ell(f,\boldsymbol{x},\boldsymbol{x}),\end{aligned}\tag{B.19}$$

首先：

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}(f)-\hat{R}_{\mathcal{S}}(f)\right)\right]\\&=\frac{1}{N_C\cdot(N_C-1)}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_in_j}\cdot T^{i,j,m,n}\right].\end{aligned}\tag{B.20}$$

给定 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$， $\mathcal{S}'=\left\{(\boldsymbol{x}'_i,y_i)\right\}_{i=1}^m$，为了便于分析，对于所有$i\leq j$将样本排序为$y_i\leq y_j$。由各样本相互独立可知：

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}(f)-\hat{R}_{\mathcal{S}}(f)\right)\right]=\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\tilde{\mathcal{S}}'}(f)-\hat{R}_{\tilde{\mathcal{S}}}(f)\right)\right]\tag{B.21}$$

其中 $\tilde{S}$ 和 $\tilde{S}'$为变换后数据集。二者从 $\mathcal{S}$ 和 $\mathcal{S}'$ 分别交换 $0\leq x\leq N$ 具有相同下标的样本$(\boldsymbol{x}_i,y_i)$ 和 $(\boldsymbol{x}'_i,y'_i)$。

此外，对任意样本独立同分布Rademacher随机变量序列，

$$\sigma=(\sigma_1^{(1)},\cdots,\sigma_{n_1}^{(1)},\cdots,\sigma_1^{(2)},\cdots,\sigma_{n_C}^{(N_C)})$$

并且对任意$\mathcal{S}$和$\mathcal{S}'$，存在一对变化数据集$\tilde{S}^{\sigma}$ 和$\tilde{S}'^{\sigma}$满足

$$\sup_{f\in\mathcal{H}}\left[\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_in_j}\cdot T^{i,j,m,n}\right]=\sup_{f\in\mathcal{H}}\left[\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f)-\hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f).\right]\tag{B.22}$$

具体地，可通过归纳法进行证明。

基. 假设 $\mathcal{S}=\{(\boldsymbol{x}_1,1),(\boldsymbol{x}_2,2)\}$ $\mathcal{S}'=\left\{(\boldsymbol{x}'_1,1),(\boldsymbol{x}'_2,2)\right\}$, $\sigma=(\sigma_1^{(1)},\sigma_1^{(2)})$。在此定义

$$T^1=T^{1,2,1,1},T^2=T^{2,1,1,1}.$$

给出以下例子：

(a) $\sigma_1^{(1)}=1,\sigma_1^{(2)}=1$。可得：

$$\begin{aligned} T^1 &= \ell(f^{(1)},\boldsymbol{x}'_1,\boldsymbol{x}'_2)-\ell(f^{(1)},\boldsymbol{x}_1,\boldsymbol{x}_2),\\ T^2 &= \ell(f^{(2)},\boldsymbol{x}'_2,\boldsymbol{x}'_1)-\ell(f^{(2)},\boldsymbol{x}_2,\boldsymbol{x}_1), \end{aligned} \tag{B.23}$$

该结论表明

$$\sup_{f\in\mathcal{H}}\left[T^1+T^2\right]=\sup_{f\in\mathcal{H}}\left[\hat{R}_{\mathcal{S}'}(f)-\hat{R}_{\mathcal{S}}(f)\right]. \tag{B.24}$$

同时也表明 $\tilde{\mathcal{S}}_\sigma=\mathcal{S}$，$\tilde{\mathcal{S}'}_\sigma=\mathcal{S}'$。

(b) $\sigma_1^{(1)}=1,\sigma_1^{(2)}=-1.$

$$\begin{aligned} T^1 &= \ell(f^{(1)},\boldsymbol{x}'_1,\boldsymbol{x}_2)-\ell(f^{(1)},\boldsymbol{x}_1,\boldsymbol{x}'_2),\\ T^2 &= \ell(f^{(2)},\boldsymbol{x}_2,\boldsymbol{x}'_1)-\ell(f^{(2)},\boldsymbol{x}'_2,\boldsymbol{x}_1), \end{aligned} \tag{B.25}$$

该结论表明

$$\sup_{f\in\mathcal{H}}\left[T^1+T^2\right]=\sup_{f\in\mathcal{H}}\left[\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f)-\hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f)\right]. \tag{B.26}$$

其中$\tilde{\mathcal{S}}_\sigma$fi $\tilde{\mathcal{S}'}_\sigma$ 通过$(\boldsymbol{x}_2,2)$ 和 $(\boldsymbol{x}'_2,2)$得到。

(c) $\sigma_1^{(1)}=-1,\sigma_1^{(2)}=1$。相应的 $\tilde{\mathcal{S}}_\sigma,\tilde{\mathcal{S}'}_\sigma$ 通过交换 $(\boldsymbol{x}_1,1)$ 和 $(\boldsymbol{x}'_1,1)$得到。

(d) $\sigma_1^{(1)}=-1,\sigma_1^{(2)}=-1$。相应的 $\tilde{\mathcal{S}}_\sigma,\tilde{\mathcal{S}'}_\sigma$ 通过交换 $\mathcal{S}$ 和 $\mathcal{S}'$得到。

通过以上讨论完成了基本情况下的证明。

递归. 给定

$$\mathcal{S}^-=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^k,\mathcal{S}^{-'}=\{(\boldsymbol{x}'_i,y_i)\}_{i=1}^k,\forall\sigma^-\in\{-1,1\}^k,$$

假设 $\mathcal{S}^-,\mathcal{S}^{-'}$ 和 σ^- 满足上述结论，且对应$\tilde{\mathcal{S}}^-_{\sigma^-}$ 和 $\tilde{\mathcal{S}}^{-'}_{\sigma^-}$存在。现在证明给定一对新的 $(\boldsymbol{x}_n,i)$, $(\boldsymbol{x}'_n,i)$ 和 $\sigma_n^{(i)}\in\{-1,1\}$ 满足

$$\mathcal{S}=\mathcal{S}^-\cup\{(\boldsymbol{x}_n,i)\},\mathcal{S}'=\mathcal{S}^{-'}\cup\{(\boldsymbol{x}'_n,i)\},\sigma=(\sigma^-,\sigma_{new}^{(i)})$$

仍然满足该结论。显然只有 $T^{i,*,n,*}$ 和 $T^{*,i,*,n}$ 和新样本相关，其中 $*$ 代表任意可能的选择。

在$\sigma_n^{(i)}=-1$的情况下，有以下结论：

(a) $\sigma_m^{(j)}=1$时：

$$T^{i,j,n,m}=\ell(f^{(i)},\boldsymbol{x}_n,\boldsymbol{x}'_m)-\ell(f^{(i)},\boldsymbol{x}'_n,\boldsymbol{x}_m).$$

(b) $\sigma_m^{(j)} = -1$时：

$$T^{i,j,n,m} = \ell(f^{(i)}, \boldsymbol{x}_n, \boldsymbol{x}_m) - \ell(f^{(i)}, \boldsymbol{x'}_m, \boldsymbol{x'}_n).$$

此外，对任意 $T^{j,i,n,m}$：

(a) $\sigma_m^{(j)} = 1$时：

$$T^{j,i,m,n} = \ell(f^{(j)}, \boldsymbol{x'}_m, \boldsymbol{x}_n) - \ell(f^{(j)}, \boldsymbol{x}_m, \boldsymbol{x'}_n).$$

(b) $\sigma_m^{(j)} = -1$时：

$$T^{i,j,n,m} = \ell(f^{(j)}, \boldsymbol{x}_m, \boldsymbol{x}_n) - \ell(f^{(j)}, \boldsymbol{x'}_m, \boldsymbol{x'}_n).$$

该结论表明

$$\tilde{\mathcal{S}}_\sigma = \tilde{\mathcal{S}}^-_{\sigma^-} \cup \{(\boldsymbol{x}'_n, i)\},\ \tilde{\mathcal{S}}'_\sigma = \tilde{\mathcal{S}}^{-'}_{\sigma^-} \cup \{(\boldsymbol{x}_n, i)\}.$$

类似地，当 $\sigma_n^{(i)} = 1$，

$$\tilde{\mathcal{S}}_\sigma = \tilde{\mathcal{S}}^-_{\sigma^-} \cup \{(\boldsymbol{x}_n, i)\},\ \tilde{\mathcal{S}}'_\sigma = \tilde{\mathcal{S}}^{-'}_{\sigma^-} \cup \{(\boldsymbol{x}'_n, i)\}.$$

另一种情况下，由 $\tilde{\mathcal{S}}_\sigma$ 和 $\tilde{\mathcal{S}}'_\sigma$ 通过交换$\mathcal{S}$ 和 $\mathcal{S}'$ 中的对应项得到。

通过一般情况中的参数和递归，完成以下证明：

$$\sup_{f\in\mathcal{H}}\left[\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_i n_j}\cdot T^{i,j,m,n}\right] = \sup_{f\in\mathcal{H}}\left[\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f) - \hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f)\right]. \tag{B.27}$$

由此可得

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_i n_j}\cdot T^{i,j,m,n}\right]\\
&= \mathop{\mathbb{E}}_{\sigma}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f) - \hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f)\right)\right]\\
&= \frac{1}{2^N}\cdot\sum_{\sigma}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\tilde{\mathcal{S}}^{\sigma'}}(f) - \hat{R}_{\tilde{\mathcal{S}}^{\sigma}}(f)\right)\right]\\
&= \frac{1}{2^N}\cdot\sum_{\sigma}\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right)\right]\\
&= \frac{2^N}{2^N}\cdot\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right)\right]\\
&= \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}'}(f) - \hat{R}_{\mathcal{S}}(f)\right)\right].
\end{aligned} \tag{B.28}$$

由此实现证明

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_i n_j}\cdot T^{i,j,m,n}\right] \leq 4\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell\circ\mathcal{H}). \tag{B.29}$$

B.5.2 MAUC$^{\downarrow}$诱导的泛化界一般形式

本节基于引理.A.3和引理.A.7给出 MAUC$^{\downarrow}$泛化界的推广结果。

再次声明定理 B.5.2 (**泛化界一般形式**). 给定数据集$\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^m$，样本由独立采样得到, 对于所有多类得分函数$f\in\mathcal{H}$，若替代损失函数$\ell$的值域包含于$[0,B]$，$\forall\delta\in(0,1)$，以下不等式至少依概率$1-\delta$成立：

$$R_\ell(f)\le\hat{R}_{\mathcal{S}}(f)+C_1\cdot\frac{\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})}{N_C(N_C-1)}+C_2\cdot\frac{B}{N_C}\cdot\xi(\boldsymbol{Y})\cdot\sqrt{\frac{\log(\frac{2}{\delta})}{N}},$$

C_1,C_2为常数，$\xi(\boldsymbol{Y})=\sqrt{\sum_{i=1}^{N_C}\frac{1}{\rho_i}},\rho_i=\frac{n_i}{N}$。

证明. 给定 $\mathcal{S}'$ 为独立于 $\mathcal{S}$的另一数据集，由于 $\hat{R}_{\mathcal{S}}(f)=\mathbb{E}_{\mathcal{S}'}\hat{R}_{\mathcal{S}'}(f)$，由Jensen's不等式可得：

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{f\in\mathcal{H}}\left(\mathop{\mathbb{E}}_{\mathcal{S}}\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}}(f)\right)\right]&=\mathop{\mathbb{E}}_{\mathcal{S}}\sup_{f\in\mathcal{H}}\mathop{\mathbb{E}}_{\mathcal{S}'}\left[\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}'}(f)\right]\\&\le\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}'}(f)\right)\right]\end{aligned}$$

由引理.A.7可得：

$$\mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f\in\mathcal{H}}\left(\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}'}(f)\right)\right]\le4\frac{1}{N_C\cdot(N_C-1)}\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell\circ\mathcal{H}).\tag{B.30}$$

给定 $\mathcal{S}$，定义 $\mathcal{S}_m=(\mathcal{S}\setminus\{(\boldsymbol{x}_m,y_m)\})\cup\{(\boldsymbol{x}'_m,y_m)\}$。现在开始推导损失函数的有界差分性质。假设 $y_m=i$，有：

$$\begin{aligned}d_i&=|\sup_{f\in\mathcal{H}}(\mathop{\mathbb{E}}_{\mathcal{S}}\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}}(f))-\sup_{f\in\mathcal{H}}(\mathbb{E}_{\mathcal{S}_i}\hat{R}_{\mathcal{S}_i}(f)-\hat{R}_{\mathcal{S}_i}(f))|\\&\le\sup_{f\in\mathcal{H}}|\hat{R}_{\mathcal{S}}(f)-\hat{R}_{\mathcal{S}_i}(f)|\\&=\frac{1}{N_C(N_C-1)}\sup_{f\in\mathcal{H}}\left[\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_in_j}|\ell(f^{(i)},\boldsymbol{x}_m,\boldsymbol{x}_n)-\ell(f^{(i)},\boldsymbol{x}'_m,\boldsymbol{x}_n)|\right.\\&\quad\left.+\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{1}{n_in_j}|\ell(f^{(j)},\boldsymbol{x}_n,\boldsymbol{x}_m)-\ell(f^{(j)},\boldsymbol{x}_n,\boldsymbol{x}'_m)|\right]\\&\le\frac{2B}{N_Cn_i}\end{aligned}$$

因此有$\sum_{i=1}^N d_i^2=4\frac{B^2}{N_C^2}\cdot\sum_{i=1}^{n_C}\frac{1}{n_i}$。根据公式.(B.30)和引理.A.3可知，固定标签$\boldsymbol{Y}$，依不小于 $1-\frac{\delta}{2}$的概率。以下不等式成立：

$$\mathop{\mathbb{E}}_{S}\hat{R}_S(f) \leq \hat{R}_S(f) + \frac{4}{N_C(N_C-1)}\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell\circ\mathcal{H}) + \frac{2B}{N_C}\xi(\boldsymbol{Y})\cdot\sqrt{\frac{\log(\frac{2}{\delta})}{2N}} \tag{B.31}$$

类似可证，固定标签 $\boldsymbol{Y}$，依不小于 $1-\frac{\delta}{2}$ 的概率，以下不等式成立：

$$\frac{\mathfrak{R}_{\mathsf{MAUC}^{\downarrow}}(\ell\circ\mathcal{H})}{N_C(N_C-1)} \leq \frac{\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},S}(\ell\circ\mathcal{H})}{N_C(N_C-1)} + \frac{2B}{N_C}\xi(\boldsymbol{Y})\cdot\sqrt{\frac{\log(\frac{2}{\delta})}{2N}}. \tag{B.32}$$

结合公式.(B.31)和公式.(B.32)，固定 $\boldsymbol{Y}$可得，依不小于$1-\delta$ 的概率：

$$\mathop{\mathbb{E}}_{S}\hat{R}_S(f) \leq \hat{R}_S(f) + \frac{4}{N_C(N_C-1)}\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},S}(\ell\circ\mathcal{H}) + \frac{10B}{N_C}\xi(\boldsymbol{Y})\cdot\sqrt{\frac{\log(\frac{2}{\delta})}{2N}} \tag{B.33}$$

根据(Agarwal 等, 2005)中的定理.8可知，式 (B.33) 在 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$上以不小于$1-\delta$的概率成立。

B.5.3　MAUC$^{\downarrow}$Rademacher复杂度的次高斯性质

定义 B.2 (次高斯随机过程). 如果对任意$\theta,\theta'\in\mathcal{T}$和所有$\lambda\in\mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda\cdot(X_\theta - X_{\theta'})\right)\right] \leq \exp\left(\frac{\lambda^2 d(\theta,\theta')^2}{2}\right). \tag{B.34}$$

则一个下标集为$\mathcal{T}$的随机过程$\theta\mapsto X_\theta$ 对于$\mathcal{T}$上的伪度量d是次高斯的。

定义 $T_f(\sigma) = \sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j} T^{i,j,m,n}$,

$$T^{i,j,m,n} = \frac{\sigma_m^{(i)}+\sigma_n^{(j)}}{2}\cdot\frac{\ell(f^{(i)}(\boldsymbol{x}_m)-f^{(i)}(\boldsymbol{x}_n))}{n_i n_j},$$

且函数 f 采样自$\mathcal{F}$。可知 $T_f(\sigma)_{f\in\mathcal{F}}$ 本质上是一个关于 Rademacher随机变量σ的随机过程。该结论通过以下引理证明。

引理 B.8. 给定输入特征集合 $\mathcal{D}_{\mathcal{X}}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$, $\mathcal{Z}_{\mathcal{D}}$ 定义为

$$\{(\boldsymbol{x},y):\boldsymbol{x}\in\mathcal{D}_{\mathcal{X}}, y\in\{1,2,\cdots,N_C\}\},$$

即 $\mathcal{D}_{\mathcal{X}}$ 和标签空间的笛卡尔积。此外，$\forall\boldsymbol{z}=(\boldsymbol{x},i)\in\mathcal{Z}_{\mathcal{D}}$, $s(\boldsymbol{z})=s^{(i)}(\boldsymbol{x})$。基于以上符号定义，若 ℓ 满足 ϕ_ℓ-Lipschitz连续，则 $\{C_G\cdot T_f(\sigma)\}_{f\in\mathcal{H}}$，其中 $f=(f^{(1)},\cdots,f^{N_C})$ 且 $f^{(i)}=\mathsf{soft}^{(i)}\circ\boldsymbol{S}$，为次高斯随机过程。具体地，

$$\mathop{\mathbb{E}}_{\sigma}\left[\exp\left(\lambda\cdot C_G\cdot\left(T_f(\sigma)-T_{\tilde{f}}(\sigma)\right)\right)\right] \leq \exp\left(\frac{\lambda^2 d_\infty(\boldsymbol{S},\tilde{\boldsymbol{s}})^2}{2}\right). \tag{B.35}$$

其中

$$C_G = \frac{1}{\phi_\ell \cdot (N_C - 1) \cdot \xi(\boldsymbol{Y}) \cdot \sqrt{\frac{1}{N}}}, \quad d_{\infty,\mathcal{S}}(\boldsymbol{S}, \tilde{\boldsymbol{s}}) = \max_{z \in \mathcal{Z}_\mathcal{D}} |\boldsymbol{S}(z) - \tilde{\boldsymbol{s}}(z)|. \tag{B.36}$$

证明. 记 σ 为Rademacher随机变量的集合：

$$\sigma = (\sigma_1^{(1)}, \cdots, \sigma_{n_1}^{(1)}, \cdots, \sigma_1^{(i)}, \cdots, \sigma_k^{(i)}, \cdots, \sigma_{n_1}^{(i)}, \cdots, \sigma_{n_{N_C}}^{(N_C)})$$

并且记 $\sigma_{\setminus(i,k)}$ 为另一个Rademacher随机变量的集合，其中各项等于 σ 除了 $\sigma_k^{(i)}$ 被替换为 $\tilde{\sigma}_k^{(i)}$，即

$$\sigma_{\setminus(i,k)} = (\sigma_1^{(1)}, \cdots, \sigma_{n_1}^{(1)}, \cdots, \sigma_1^{(i)}, \cdots, \tilde{\sigma}_k^{(i)}, \cdots, \sigma_{n_1}^{(i)}, \cdots, \sigma_{n_{N_C}}^{(N_C)}).$$

对任意(i,k)有以下有限差分性质:

$$\begin{aligned}
diff_i &= \left|\left(T_f(\sigma) - T_{\tilde{f}}(\sigma)\right) - \left(T_f(\sigma_{\setminus(i,k)}) - T_{\tilde{f}}(\sigma_{\setminus(i,k)})\right)\right| \\
&= \left|\left(\frac{\sigma_k^{(i)} - \tilde{\sigma}_k^{(i)}}{2}\right) \cdot \left[\sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \cdot \frac{1}{n_i n_j} \left(\ell(f^{(i)}, \boldsymbol{x}_m, \boldsymbol{x}_n) - \ell(\tilde{f}^{(i)}, \boldsymbol{x}_m, \boldsymbol{x}_n)\right)\right.\right. \\
&\quad \left.\left. + \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{1}{n_i n_j} \left(\ell(f^{(j)}, \boldsymbol{x}_n, \boldsymbol{x}_m) - \ell(\tilde{f}^{(j)}, \boldsymbol{x}_n, \boldsymbol{x}_m)\right)\right]\right| \\
&\leq \frac{2 \cdot (N_C - 1)}{n_i} \max_{i,\, \boldsymbol{x}_m \in \mathcal{N}_i,\, \boldsymbol{x}_n \notin \mathcal{N}_i} \left|\ell(f^{(i)}, \boldsymbol{x}_m, \boldsymbol{x}_n) - \ell(\tilde{f}^{(i)}, \boldsymbol{x}_m, \boldsymbol{x}_n)\right| \\
&\leq \frac{4 \cdot (N_C - 1)}{n_i} \cdot \phi_\ell \cdot \max_{i,\, \boldsymbol{x} \in \mathcal{D}_\mathcal{X}} \left|f^{(i)}(\boldsymbol{x}) - \tilde{f}^{(i)}(\boldsymbol{x})\right| \\
&= \frac{4 \cdot (N_C - 1)}{n_i} \cdot \phi_\ell \cdot \max_{i,\, \boldsymbol{x} \in \mathcal{D}_\mathcal{X}} \left|\int_0^1 \left\langle \nabla \text{soft}^{(i)}\left(\tau \cdot \tilde{\boldsymbol{s}}(\boldsymbol{x}) + (1-\tau) \cdot \boldsymbol{S}(\boldsymbol{x})\right), \left(\boldsymbol{S}(\boldsymbol{x}) - \tilde{\boldsymbol{s}}(\boldsymbol{x})\right)\right\rangle d\tau\right| \\
&\leq \frac{4 \cdot (N_C - 1)}{n_i} \cdot \phi_\ell \cdot \left(\sup_{\boldsymbol{x}} ||\nabla \text{soft}^{(i)}(\boldsymbol{x})||_1\right) \cdot \max_{z \in \mathcal{Z}_\mathcal{D}} |\boldsymbol{S}(z) - \tilde{\boldsymbol{s}}(z)| \\
&\leq \frac{2 \cdot (N_C - 1)}{n_i} \cdot \phi_\ell \cdot \max_{z \in \mathcal{Z}_\mathcal{D}} |\boldsymbol{S}(z) - \tilde{\boldsymbol{s}}(z)|.
\end{aligned}$$

显然，$C_G \cdot \left(T_f(\sigma) - T_{\tilde{f}}(\sigma)\right)$满足有限差分性质 (定义.B.1)。根据引理.B.1，可选择v：

$$v = \frac{1}{4} \sum_{i=1}^{N} C_G^2 \cdot diff_i^2 = \max_{z \in \mathcal{Z}_\mathcal{D}} |\boldsymbol{s}(z) - \tilde{\boldsymbol{s}}(z)|^2$$

基于引理.B.1即完成证明。 □

B.5.4　Chaining界

再次声明定理 *B.5.4* (MAUC$^{\downarrow}$ Rademacher复杂度的chaining界). 假设得分函数$s^{(i)}$将$\mathcal{X}$映射至有界区间$[-R_s, R_s]$，$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H})$具有以下性质：

(a) 对于任意单减序列$\{\epsilon_k\}_{k=1}^{\infty}$，若$\lim_{k\to\infty}\epsilon_k = 0$且$\epsilon_0 \geq R_s$，则：

$$\begin{aligned}\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H}) &\leq N_C \cdot (N_C - 1) \cdot \phi_\ell \cdot \epsilon_K \\ &+ 6 \cdot \sum_{k=1}^{K} \epsilon_k \phi_\ell (N_C - 1) \cdot \xi(\boldsymbol{Y}) \sqrt{\frac{\log(\mathfrak{C}(\epsilon_k, \mathcal{F}, d_{\infty,\mathcal{S}}))}{N}}\end{aligned}$$

(b) 存在一个常数C，满足：

$$\begin{aligned}\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H}) \leq C\phi_\ell \inf_{R_s \geq \alpha \geq 0} \Bigg(& N_C(N_C - 1)\alpha \\ &+ (N_C - 1) \cdot \xi(\boldsymbol{Y}) \cdot \int_{\alpha}^{R_s} \sqrt{\frac{\log(\mathfrak{C}(\epsilon, \mathcal{F}, d_{\infty,\mathcal{S}}))}{N}} d\epsilon \Bigg)\end{aligned}$$

证明. 首先：

$$\begin{aligned}2C_G \hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H}) &= C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f \in \mathcal{H}} T_f(\sigma) + \sup_{\tilde{f} \in \mathcal{H}} T_{\tilde{f}}(\sigma) \right] \\ &= C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f \in \mathcal{H}} T_f(\sigma) + \sup_{\tilde{f} \in \mathcal{H}} T_{\tilde{f}}(-\sigma) \right] \\ &= C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f \in \mathcal{H}} T_f(\sigma) + \sup_{\tilde{f} \in \mathcal{H}} -T_{\tilde{f}}(\sigma) \right] \\ &= C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f, \tilde{f} \in \mathcal{H}} \left(T_f(\sigma) - T_{\tilde{f}}(\sigma) \right) \right]\end{aligned} \tag{B.37}$$

定义$\hat{\mathcal{H}}$为 $\mathcal{H}$ 的 ϵ-覆盖，且有伪度量 $d_{\infty,\mathcal{N}}$。选择 $\hat{f}, \hat{\tilde{f}} \in \hat{\mathcal{H}}$，满足 $d_{\infty,\mathcal{N}}(f, \hat{f}) \leq \epsilon$ 和 $d_{\infty,\mathcal{N}}(\tilde{f}, \hat{\tilde{f}}) \leq \epsilon$。可得以下结果：

$$\begin{aligned}&C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f, \tilde{f} \in \mathcal{H}} \left(T_f(\sigma) - T_{\tilde{f}}(\sigma) \right) \right] \\ &= C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f, \tilde{f} \in \mathcal{H}} \left(T_f(\sigma) - T_{\hat{f}}(\sigma) \right) + \left(T_{\hat{f}}(\sigma) - T_{\hat{\tilde{f}}}(\sigma) \right) + \left(T_{\hat{\tilde{f}}}(\sigma) - T_{\tilde{f}}(\sigma) \right) \right] \\ &\leq 2C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{d_{\infty,\mathcal{N}}(\boldsymbol{s}, \hat{\boldsymbol{s}}) \leq \epsilon} \left(T_f(\sigma) - T_{\hat{f}}(\sigma) \right) \right] + C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\hat{\boldsymbol{s}}, \hat{\tilde{\boldsymbol{s}}} \in \hat{\mathcal{H}}_\epsilon} \left(T_{\hat{f}}(\sigma) - T_{\hat{\tilde{f}}}(\sigma) \right) \right]\end{aligned} \tag{B.38}$$

进一步地，令 $\hat{\mathcal{H}}_k$为$\mathcal{H}$的 ϵ_k-覆盖。对任意k，选择$\hat{\boldsymbol{s}}_k$，$\hat{\tilde{\boldsymbol{s}}}_k$ 使其满足 $d_{\infty,\mathcal{S}}(\hat{\boldsymbol{s}}_k, \boldsymbol{s}) \leq \epsilon_k$ 和 $d_{\infty,\mathcal{S}}(\hat{\tilde{\boldsymbol{s}}}_k, \tilde{\boldsymbol{s}}) \leq \epsilon_k$。具体地，$\epsilon_K = \epsilon$，$\epsilon_0 \geq R_s$。基于此，选择 $\hat{\boldsymbol{s}}_0 = \hat{\tilde{\boldsymbol{s}}}_0$，$\epsilon_{k+1} = \frac{1}{2}\epsilon_k$，

$\hat{f}_K = \hat{f}$，和 $\hat{\tilde{f}}_K = \hat{\tilde{f}}$。另有 $\hat{f}_k = \text{soft} \circ \hat{\boldsymbol{s}}_k$ 和 $\hat{\tilde{f}}_k = \text{soft} \circ \hat{\tilde{\boldsymbol{s}}}_k$。则对所有 $\boldsymbol{s}, \tilde{\boldsymbol{s}} \in \hat{\mathcal{H}}_\epsilon$，可将 $T_{\hat{f}}(\sigma)$ 表示为

$$T_{\hat{f}}(\sigma) = T_{\hat{f}_K}(\sigma) = T_{\hat{f}_0}(\sigma) + \sum_{i=1}^{K} \left(T_{\hat{f}_k}(\sigma) - T_{\hat{f}_{k-1}}(\sigma) \right)$$

和

$$T_{\hat{\tilde{f}}}(\sigma) = T_{\hat{\tilde{f}}_K}(\sigma) = T_{\hat{\tilde{f}}_0}(\sigma) + \sum_{i=1}^{K} \left(T_{\hat{\tilde{f}}_k}(\sigma) - T_{\hat{\tilde{f}}_{k-1}}(\sigma) \right).$$

由此可得：

$$C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\hat{\boldsymbol{s}}, \hat{\tilde{\boldsymbol{s}}} \in \hat{\mathcal{H}}_\epsilon} \left(T_{\hat{f}}(\sigma) - T_{\hat{\tilde{f}}}(\sigma) \right) \right] \leq 2 \cdot C_G \cdot \sum_{i=1}^{K} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\substack{\boldsymbol{S}_k \in \hat{\mathcal{H}}_k, \hat{\boldsymbol{S}}_{k-1} \in \hat{\mathcal{H}}_{k-1} \\ d_{\infty,\mathcal{S}}(\hat{\boldsymbol{S}}_k, \hat{\boldsymbol{S}}_{k-1}) \leq 3\epsilon_k}} \left(T_{\hat{f}_k}(\sigma) - T_{\hat{f}_{k-1}}(\sigma) \right) \right]. \tag{B.39}$$

根据极大值不等式 (引理.B.4.1.2)可得：

$$\begin{aligned} & C_G \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\substack{\boldsymbol{S}_k \in \hat{\mathcal{H}}_k, \hat{\boldsymbol{S}}_{k-1} \in \hat{\mathcal{H}}_{k-1} \\ d_{\infty,\mathcal{S}}(\hat{\boldsymbol{S}}_k, \hat{\boldsymbol{S}}_{k-1}) \leq 3\epsilon_k}} \left(T_{\hat{f}_k}(\sigma) - T_{\hat{f}_{k-1}}(\sigma) \right) \right] \\ & \leq 3\epsilon_k \sqrt{2 \log |\hat{\mathcal{H}}_k| \cdot |\hat{\mathcal{H}}_{k-1}|} \leq 6\epsilon_k \sqrt{\log(\mathfrak{C}(\epsilon_k, \mathcal{F}, d_{\infty,\mathcal{S}}))}. \end{aligned} \tag{B.40}$$

对引理.B.8进行类似的推导可得：

$$\mathop{\mathbb{E}}_{\sigma} \left[\sup_{d_{\infty,\mathcal{N}}(\boldsymbol{s}, \hat{\boldsymbol{s}}) \leq \epsilon_K} \left(T_f(\sigma) - T_{\hat{f}}(\sigma) \right) \right] \leq N_C \cdot (N_C - 1) \cdot \phi_\ell \cdot \epsilon_K \tag{B.41}$$

综上可得：

$$\begin{aligned} \hat{\mathfrak{R}}_{\text{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathcal{H}) &= \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f, \tilde{f} \in \mathcal{H}} \left(T_f(\sigma) - T_{\tilde{f}}(\sigma) \right) \right] \\ &\leq N_C \cdot (N_C - 1) \cdot \phi_\ell \cdot \epsilon_K + \left(\frac{1}{C_G} \right) 6 \sum_{i=1} \epsilon_k \sqrt{\log(\mathfrak{C}(\epsilon_k, \mathcal{F}, d_{\infty,\mathcal{S}}))} \\ &\leq N_C \cdot (N_C - 1) \cdot \phi_\ell \cdot \epsilon_K + 6 \sum_{i=1}^{K} \epsilon_k \phi_\ell \cdot (N_C - 1) \xi(\boldsymbol{Y}) \sqrt{\frac{\log(\mathfrak{C}(\epsilon_k, \mathcal{F}, d_{\infty,\mathcal{S}}))}{N}} \end{aligned}$$

(a)证明完毕。

基于(a)的结果证明(b)。首先令 $\epsilon_k = 2(\epsilon_k - \epsilon_{k+1})$，且 $\log(\mathfrak{C}(\epsilon, \mathcal{F}, d_{\infty,\mathcal{S}})$ 对于 ϵ 非递增。由此可得：

$$\begin{aligned} & N_C \cdot (N_C - 1) \cdot \phi_\ell \cdot \epsilon_K + 6 \sum_{i=1}^{K} \epsilon_k \phi_\ell \cdot (N_C - 1) \cdot \xi(\boldsymbol{Y}) \sqrt{\frac{\log(\mathfrak{C}(\epsilon_k, \mathcal{F}, d_{\infty,\mathcal{S}}))}{N}} \\ & \leq 2N_C \cdot (N_C - 1) \cdot \phi_\ell \cdot \epsilon_{K+1} + 12 \sum_{i=1}^{K} (\epsilon_k - \epsilon_{k+1}) \cdot (N_C - 1) \phi_\ell \cdot \xi(\boldsymbol{Y}) \sqrt{\frac{\log(\mathfrak{C}(\epsilon_k, \mathcal{F}, d_{\infty,\mathcal{S}}))}{N}} \end{aligned}$$

进一步可得：

$$\begin{aligned}\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H}) \leq 12\phi_\ell\,(N_C\cdot(N_C-1)\cdot\epsilon_{K+1}\\ +\sum_{i=1}^{K}(\epsilon_k-\epsilon_{k+1})\cdot(N_C-1)\cdot\xi(\boldsymbol{Y})\sqrt{\frac{\log(\mathfrak{C}(\epsilon_k,\mathcal{F},d_{\infty,\mathcal{S}}))}{N}}\Bigg)\\ \leq 12\phi_\ell\Bigg(N_C\cdot(N_C-1)\cdot\epsilon_{K+1}+\int_{\epsilon_{K+1}}^{R_s}(N_C-1)\xi(\boldsymbol{Y})\sqrt{\frac{\log(\mathfrak{C}(\epsilon_k,\mathcal{F},d_{\infty,\mathcal{S}}))}{N}}d\epsilon\\ +\int_{R_s}^{\epsilon_0}(N_C-1)\xi(\boldsymbol{Y})\sqrt{\frac{\log(\mathfrak{C}(\epsilon_k,\mathcal{F},d_{\infty,\mathcal{S}}))}{N}}d\epsilon\Bigg)\\ =12\phi_\ell\left(N_C\cdot(N_C-1)\cdot\epsilon_{K+1}+\int_{\epsilon_{K+1}}^{R_s}\xi(\boldsymbol{Y})\sqrt{\frac{\log(\mathfrak{C}(\epsilon_k,\mathcal{F},d_{\infty,\mathcal{S}}))}{N}}d\epsilon\right)\end{aligned}$$

由此可知常数应选择 $C=12$。令 $\alpha=\epsilon_{K+1}$。通过恰当选择K 和ϵ_0，并令 $\alpha=\frac{\epsilon_0}{2^{K+1}}$，可知对于所有$\alpha\in[0,R_s]$不等式都成立。证明完毕。 □

B.5.5　关键引理

B.5.5.1　一类全连接网络的MAUC$^{\downarrow}$Rademacher复杂度

引理 B.9.

$$\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathrm{soft}\circ\mathcal{H}_\gamma^{DNN,nh})\leq\frac{\sqrt{2}}{2}\phi_\ell R_{\mathcal{X}}\gamma N_C\xi(\boldsymbol{Y})\left(\sqrt{\frac{L\log 2(N_C-1)}{N}}+\sqrt{\frac{N_C(N_C-1)}{N}}\right)$$
$$+\frac{\sqrt{2}}{2}\phi_\ell N_C\chi(\boldsymbol{Y})\sqrt{\frac{1}{N}}$$

其中 $\chi(\boldsymbol{Y})=\sqrt{\sum_{i=1}^{N_C}\sum_{j\neq i}\frac{1}{\rho_i,\rho_j}}$, $\xi(\boldsymbol{Y})=\sqrt{\sum_{i=1}^{N_C}\frac{1}{\rho_i}}$, $\rho_i=\frac{n_i}{N}$.

证明. 由压缩引理，可得到以下泛化界：

$$\begin{aligned}&\hat{\mathfrak{R}}_{\mathrm{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathrm{soft}\circ\mathcal{H}_\gamma^{DNN,nh})\\ &\leq\phi_\ell\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\sup_{g\in\mathrm{soft}\circ\mathcal{H}_\gamma^{DNN,nh}}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\frac{\sigma_m^{(i)}}{2}\cdot\frac{1}{n_in_j}\cdot(g^{(i)}(\boldsymbol{x}_m)-g^{(i)}(\boldsymbol{x}_n))\right]\end{aligned}$$

$$+\phi_\ell\sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\mathop{\mathbb{E}}_{\sigma^{(j)}}\left[\sup_{g\in\mathrm{soft}\circ\mathcal{H}_\gamma^{DNN,nh}}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{\sigma_n^{(j)}}{2}\cdot\frac{1}{n_in_j}\cdot(g^{(i)}(\boldsymbol{x}_m)-g^{(i)}(\boldsymbol{x}_n))\right]$$

由上确界的次可加性可得：

$$\begin{aligned}
&\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathsf{soft} \circ \mathcal{H}_{\gamma}^{DNN,nh}) \\
&\leq \underbrace{\phi_{\ell} \sum_{i=1}^{N_C} (N_C - 1) \cdot \mathop{\mathbb{E}}_{\sigma^{(i)}} \left[\sup_{g \in \mathsf{soft} \circ \mathcal{H}_{\gamma}^{DNN,nh}} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \frac{\sigma_m^{(i)}}{2} \cdot \frac{1}{n_i} \cdot (g^{(i)}(\boldsymbol{x}_m)) \right]}_{(a)} \\
&\quad + \underbrace{\phi_{\ell} \sum_{i=1}^{N_C} \mathop{\mathbb{E}}_{\sigma^{(j)}} \left[\sup_{g \in \mathsf{soft} \circ \mathcal{H}_{\gamma}^{DNN,nh}} \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{\sigma_n^{(j)}}{2} \cdot \frac{1}{n_j} \cdot (g^{(i)}(\boldsymbol{x}_n)) \right]}_{(b)} \\
&\quad + \underbrace{\phi_{\ell} \sum_{i=1}^{N_C} \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \mathop{\mathbb{E}}_{\sigma^{(i)}} \left[\sup_{g \in \mathsf{soft} \circ \mathcal{H}_{\gamma}^{DNN,nh}} (g^{(i)}(\boldsymbol{x}_n)) \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \frac{\sigma_m^{(i)}}{2} \cdot \frac{1}{n_i n_j} \right]}_{(c)} \\
&\quad + \underbrace{\phi_{\ell} \sum_{i=1}^{N_C} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \mathop{\mathbb{E}}_{\sigma^{(j)}} \sup_{g \in \mathsf{soft} \circ \mathcal{H}_{\gamma}^{DNN,nh}} \left[(g^{(i)}(\boldsymbol{x}_m)) \cdot \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \frac{\sigma_n^{(j)}}{2} \cdot \frac{1}{n_i n_j} \right]}_{(d)}
\end{aligned} \tag{B.42}$$

首先推导 $(a) + (b)$的界。

根据引理.B.4和引理.B.6可知：

$$\begin{aligned}
(a) + (b) \leq \phi_{\ell} \sum_{i=1}^{N_C} (N_C - 1) \cdot \mathop{\mathbb{E}}_{\sigma^{(i)}} \left[\sup_{f \in \mathcal{H}_{\gamma,R_S,n_h}^{DNN}} \sum_{\boldsymbol{x}_m \in \mathcal{N}_i} \sum_{c=1}^{N_C} \frac{\sigma_{m,c}^{(i)}}{2} \cdot \frac{1}{n_i} \cdot (f^{(c)}(\boldsymbol{x}_m)) \right] \\
+ \phi_{\ell} \sum_{i=1}^{N_C} \mathop{\mathbb{E}}_{\sigma^{(j)}} \left[\sup_{f \in \mathcal{H}_{\gamma,R_S,n_h}^{DNN}} \sum_{j \neq i} \sum_{\boldsymbol{x}_n \in \mathcal{N}_j} \sum_{c=1}^{N_C} \frac{\sigma_{n,c}^{(j)}}{2} \cdot \frac{1}{n_j} \cdot (f^{(c)}(\boldsymbol{x}_n)) \right]
\end{aligned}$$

递归地将引理.B.5应用于神经网络的各层：

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\sup_{f\in\mathcal{H}^{DNN}_{\gamma,R_s,n_h}}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_in_j}\cdot(f^{(c)}(\boldsymbol{x}_m))\right]\\
&\quad\le\frac{1}{\lambda}\log\left(\exp\left(\lambda\cdot\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\sup_{f\in\mathcal{H}^{DNN}_{\gamma,R_s,n_h}}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_in_j}\cdot(f^{(c)}(\boldsymbol{x}_m))\right]\right)\right)\\
&\quad\le\frac{1}{\lambda}\log\left(\left(\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\sup_{f\in\mathcal{H}^{DNN}_{\gamma,R_s,n_h}}\exp\left(\lambda\cdot\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_in_j}\cdot(f^{(c)}(\boldsymbol{x}_m))\right)\right]\right)\right)\\
&\quad\le\frac{1}{\lambda}\log\left(2^L\cdot\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\exp\left(\lambda\gamma\left\|\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_in_j}\cdot\boldsymbol{x}_m\right\|\right)\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma^{(j)}}\left[\sup_{f\in\mathcal{H}^{DNN}_{\gamma,R_s,n_h}}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\sum_{c=1}^{N_C}\frac{\sigma^{(j)}_{n,c}}{2}\cdot\frac{1}{n_in_j}\cdot(f^{(c)}(\boldsymbol{x}_n))\right]\\
&\quad\le\frac{1}{\lambda}\log\left(2^L\cdot\mathop{\mathbb{E}}_{\sigma^{(j)}}\left[\exp\left(\lambda\gamma\left\|\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\sum_{c=1}^{N_C}\frac{\sigma^{(j)}_{n,c}}{2}\cdot\frac{1}{n_in_j}\cdot\boldsymbol{x}_n\right\|\right)\right]\right).
\end{aligned}
$$

该结论表明：

$$
\begin{aligned}
(a+b)\le\phi_\ell\sum_{i=1}^{N_C}\frac{1}{\lambda}\log\Bigg(2^L\cdot\mathop{\mathbb{E}}_{\sigma}\Bigg[\exp\Bigg(&\lambda\gamma(N_C-1)\left\|\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_i}\cdot\boldsymbol{x}_m\right\|\\
&+\lambda\gamma\left\|\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\sum_{c=1}^{N_C}\frac{\sigma^{(j)}_{n,c}}{2}\cdot\frac{1}{n_j}\cdot\boldsymbol{x}_n\right\|\Bigg)\Bigg]\Bigg),
\end{aligned}
$$

原因在于当$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 独立时，具有属性$\mathbb{E}_{\boldsymbol{a},\boldsymbol{b}}(f(\boldsymbol{a})\cdot g(\boldsymbol{a}))=\mathbb{E}_{\boldsymbol{a}}(f(\boldsymbol{a}))\cdot\mathbb{E}_{\boldsymbol{b}}(g(\boldsymbol{b}))$，

记

$$
\begin{aligned}
\boldsymbol{Z}_i=\gamma(N_C-1)&\left\|\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_i}\cdot\boldsymbol{x}_m\right\|\\
&+\gamma\left\|\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\sum_{c=1}^{N_C}\frac{\sigma^{(j)}_{n,c}}{2}\cdot\frac{1}{n_j}\cdot\boldsymbol{x}_n\right\|
\end{aligned}
\tag{B.43}
$$

为一个和Rademacher随机变量相关的随机变量。可得到：

$$
(a)+(b)\le\phi_\ell\left(\sum_{i=1}^{N_C}\frac{L\log 2+\log(\mathbb{E}_\sigma\exp(\lambda(\boldsymbol{Z}_i-\mathbb{E}(\boldsymbol{Z}_i))))}{\lambda}+\sum_{i=1}^{N_C}\mathop{\mathbb{E}}_{\sigma}(\boldsymbol{Z}_i)\right).
$$

可得

$$
\sum_{i=1}^{N_C}\mathop{\mathbb{E}}_{\sigma}(\boldsymbol{Z}_i)\le\frac{\sqrt{2}}{2}\gamma\cdot R_{\mathcal{X}}\cdot(N_C)^{3/2}\left((N_C-1)\sum_{i=1}^{N_C}\cdot\sum_{j\neq i}\frac{1}{\rho_i}\right)^{1/2}\cdot\frac{1}{\sqrt{N}}.
\tag{B.44}
$$

利用$\boldsymbol{Z}_i$的有限差分不等式可得：

$$\mathop{\mathbb{E}}_{\sigma}\exp\left(\lambda(\boldsymbol{Z}_i-\mathbb{E}(\boldsymbol{Z}_i)\right)\leq\exp\left(\frac{\lambda^2 v_i}{2}\right),\ \ v_i\leq\frac{N_C}{4}\frac{(N_C-1)^2}{n_i}\gamma^2R_{\mathcal{X}}^2+\frac{N_C}{4}\sum_{j\neq i}\frac{\gamma^2R_{\mathcal{X}}^2}{n_j},\tag{B.45}$$

由此可得：

$$\left(\sum_{i=1}^{N_C}\frac{L\log 2+\log(\mathbb{E}_\sigma\exp(\lambda(\boldsymbol{Z}_i-\mathbb{E}(\boldsymbol{Z}_i))))}{\lambda}\right)\leq\left(\sum_{i=1}^{N_C}\frac{L\log 2+(\frac{\lambda^2 v_i}{2})}{\lambda}\right)\tag{B.46}$$

通过选择 $\lambda=\sqrt{\frac{2N_CL\log 2}{\sum_{i=1}^{N_C}v_i}}$，可得最优界为：

$$\begin{aligned}\sum_{i=1}^{N_C}\frac{L\log 2+\log(\mathbb{E}_\sigma\exp(\lambda(\boldsymbol{Z}_i-\mathbb{E}(\boldsymbol{Z}_i))))}{\lambda}&\leq\sqrt{2N_CL\log 2\cdot\sum_{i=1}^{N_C}v_i}\\&=R\cdot\gamma\cdot\sqrt{\frac{L\log 2}{2N}}\cdot\sqrt{\sum_{i=1}^{N_C}\frac{N_C-1}{\rho_i}}.\end{aligned}\tag{B.47}$$

综上：

$$(a)+(b)\leq\frac{\sqrt{2}}{2}\phi_\ell R_{\mathcal{X}}\gamma\cdot(N_C(N_C-1))^{1/2}\xi(\boldsymbol{Y})\left(\sqrt{\frac{N_CL\log 2}{N}}+N_C\sqrt{\frac{1}{N}}\right)\tag{B.48}$$

对于 (c) 和 (d)：

$$(c)\leq\phi_\ell\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\left(\sup_{f\in\mathcal{H}_{\gamma,R_s,n_h}^{DNN}}\left|f^{(i)}(\boldsymbol{x}_n)\right|\right)\cdot\left|\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\frac{\sigma_m^{(i)}}{2}\cdot\frac{1}{n_in_j}\right|\right]\tag{B.49}$$

$$\overset{(*)}{\leq}\phi_\ell\cdot\sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\left|\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\frac{\sigma_m^{(i)}}{2}\cdot\frac{1}{n_in_j}\right|\right]$$

类似地：

$$(d)\overset{(**)}{\leq}\phi_\ell\sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\mathop{\mathbb{E}}_{\sigma^{(j)}}\left[\left|\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\frac{\sigma_m^{(j)}}{2}\cdot\frac{1}{n_in_j}\right|\right],\tag{B.50}$$

其中 (*) 和 (**) 由softmax输出不大于1得到，即

$$\sup_{f\in\mathcal{H}^{DNN}_{\gamma,R_s,n_h}} |f^{(i)}_{W,L}(\boldsymbol{x}_n)| \le 1 \text{ 且 } \sup_{f\in\mathcal{H}^{DNN}_{\gamma,R_s,n_h}} |f^{(i)}_{W,L}(\boldsymbol{x}_m)| \le 1.$$

基于公式.(B.49)和公式.(B.50)，证明完毕。

基于以上分析可知：

$$\begin{aligned}(c)+(d) \le{}& \phi_\ell \cdot \sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j} \mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\left|\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_i n_j}\right|\right] \\ &+ \phi_\ell \cdot \sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i} \mathop{\mathbb{E}}_{\sigma^{(j)}}\left[\left|\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{n,c}}{2}\cdot\frac{1}{n_i n_j}\right|\right] \\ \le{}& \phi_\ell \cdot \sum_{i=1}^{N_C}\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j} \sqrt{\mathop{\mathbb{E}}_{\sigma^{(i)}}\left[\left(\sum_{\boldsymbol{x}_m\in\mathcal{N}_i}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{m,c}}{2}\cdot\frac{1}{n_i n_j}\right)^2\right]} \\ &+ \phi_\ell \cdot \sum_{i=1}^{N_C}\sum_{\boldsymbol{x}_m\in\mathcal{N}_i} \sqrt{\mathop{\mathbb{E}}_{\sigma^{(j)}}\left[\left(\sum_{j\neq i}\sum_{\boldsymbol{x}_n\in\mathcal{N}_j}\sum_{c=1}^{N_C}\frac{\sigma^{(i)}_{n,c}}{2}\cdot\frac{1}{n_i n_j}\right)^2\right]}\end{aligned} \tag{B.51}$$

可得：

$$(c)+(d) \le \frac{\sqrt{2}}{2}\cdot\phi_\ell\cdot\chi(\boldsymbol{Y})\cdot N_C\sqrt{\frac{1}{N}} \tag{B.52}$$

此外，可通过chaining技术给出另一种上界。

引理 B.10. 基于定理.3.7的设定。

对所有$f \in \mathsf{soft}\circ\mathcal{H}^{DNN}_{\gamma,R_s,n_h}$，以下不等式依不低于$1-\delta$的概率成立：

$$R_{surr}(f) \le \hat{R}_{\mathcal{S}}(f) + \mathcal{I}_{DNN,2}\cdot\sqrt{\frac{1}{N}}$$

其中 $\xi(\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{N_C}\frac{1}{\rho_i}}$，$\rho_i = \dfrac{n_i}{N}$，

$$\begin{aligned}\mathcal{I}_{DNN,2} =& C_1\phi_\ell\left(\frac{2^9}{N_C}\cdot\xi(\boldsymbol{Y})\cdot\log^{3/2}(K\cdot N\cdot N_C)\cdot\gamma\cdot R_{\mathcal{X}}\cdot(\sqrt{2\log(2)L}+1)+1\right) \\ &+ C_2\frac{B\cdot\sqrt{\log(\frac{2}{\delta})}\cdot\xi(\boldsymbol{Y})}{N_C}\end{aligned}$$

与定理.B.5.2相同， C_1, C_2 为常数，$K = e\cdot R_s$。

证明. 由定理.3.6以及(Golowich 等, 2018)中定理.1

$$\hat{\mathfrak{R}}_{N\cdot N_C}(\Pi\circ\mathcal{F}) \leq \frac{R_{\mathcal{X}}\gamma\cdot(\sqrt{2\log(2)L}+1)}{\sqrt{N_C\cdot N}},$$

即可完成证明。

B.5.5.2 一类深度卷积网络的MAUC$^{\downarrow}$ Rademacher复杂度

引理 B.11. 将假设类表示为：

$$\text{soft}\circ\mathcal{F}_{\beta,\nu} = \left\{\boldsymbol{g}(\boldsymbol{x}) = \text{soft}(\boldsymbol{s}_{\boldsymbol{P}}(\boldsymbol{x})) :\ \boldsymbol{s}_{\boldsymbol{P}}\in\mathcal{F}_{\beta,\nu}\right\},$$

$$\mathcal{F}_{\beta,\nu} = \{\boldsymbol{s}_{\boldsymbol{P}} : \mathbb{R}^{N_{N_L-1}}\to\mathbb{R}^{N_C} |\ \boldsymbol{P}\in\mathcal{P}_{\beta,\nu}, Range(\boldsymbol{s}_P)\subseteq[-R_s,R_s]^{N_C}\}$$

此外，定义 $\tilde{N} = \dfrac{1}{\sum_{i=1}^{N_C}\frac{1}{n_i}}$ 。假设$\sup_{x\in\mathcal{X}}||vec(\boldsymbol{x})||\leq R_{\mathcal{X}}$ 以及

$$R_s > 1/\min\left\{\sqrt{N}, \frac{\xi(\boldsymbol{Y})}{N_C}\cdot\sqrt{N_{par}\left(\nu N_L+\beta+\log(3R_{\mathcal{X}}\cdot\beta\cdot N)\right)}\right\},$$

，可得：

$$\begin{aligned}&\hat{\mathfrak{R}}_{\text{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\text{soft}\circ\mathcal{F}_{\beta,\nu})\\ &\leq\tilde{C}\left(\phi_\ell\cdot(N_C-1)\cdot R_s\cdot\xi(\boldsymbol{Y})\cdot\sqrt{\frac{N_{par}\left(\nu N_L+\beta+\log\left(3\beta R_{\mathcal{X}}N\right)\right)}{N}}\right),\end{aligned}\tag{B.53}$$

其中 $N_L = N_{conv}+N_{conn}$, N_{par} 为参数总量，$\tilde{C}$为常数。

证明. 由(Long 等, 2020, 引理.3.4)可得

$$\max_{z\in\mathcal{Z}_{\mathcal{D}}}|\boldsymbol{s}_{\boldsymbol{P}}(z)-\tilde{\boldsymbol{s}}_{\tilde{\boldsymbol{P}}}(z)|\leq\sup_{x\in\mathcal{X}}||\boldsymbol{s}_{\boldsymbol{P}}(\boldsymbol{x})-\tilde{\boldsymbol{s}}_{\tilde{\boldsymbol{P}}}(\boldsymbol{x})||\leq R_{\mathcal{X}}\cdot\beta\cdot\left(1+\nu+\frac{\beta}{N_L}\right)^{N_L}\cdot d_{NN}(\boldsymbol{P},\tilde{\boldsymbol{P}}).$$

进一步地，根据(Long 等, 2020, Lem.A.8)可得：

$$\log(\mathfrak{C}(\epsilon,\mathcal{F}_{\beta,\nu},d_{\infty,\mathcal{S}})\leq N_{par}\cdot\log\left(\frac{3C_L}{\epsilon}\right),$$

其中 $C_L = R_{\mathcal{X}}\cdot\beta\exp(\nu N_L+\beta)$。根据定理.B.5.4-b)可得：

$$\begin{aligned}&\hat{\mathfrak{R}}_{\text{MAUC}^{\downarrow},\mathcal{S}}(\ell\circ\mathcal{H})\leq C\cdot\inf_{R_s\geq\alpha\geq0}\\ &\left(N_C\cdot(N_C-1)\cdot\phi_\ell\cdot\alpha+\phi_\ell\cdot(N_C-1)\cdot\int_\alpha^{R_s}\sqrt{\frac{N_{par}\left(\nu N_L+\beta+\log(3\beta R_{\mathcal{X}}/\epsilon)\right)}{\tilde{N}}}d\epsilon\right)\end{aligned}$$

由于 $R_s > 1/\sqrt{N}$，可以选择$\alpha = \sqrt{\frac{1}{N}}$。该结果由以下不等式得出：

$$\int_{\alpha}^{R_s} \sqrt{\frac{N_{par}\left(\nu N_L + \beta + \log(3\beta R_{\mathcal{X}}/\epsilon)\right)}{\tilde{N}}} d\epsilon \leq R_s \cdot \xi(\boldsymbol{Y}) \cdot \sqrt{\frac{N_{par}\left(\nu N_L + \beta + \log\left(3\beta R_{\mathcal{X}} N\right)\right)}{N}}.$$

基于 $R_s > 1/\left(\frac{\xi(\boldsymbol{Y})}{N_C} \cdot \sqrt{N_{par}\left(\nu N_L + \beta + \log(3R_{\mathcal{X}} \cdot \beta \cdot N)\right)}\right)$, 可得：

$$\hat{\mathfrak{R}}_{\mathsf{MAUC}^{\downarrow},\mathcal{S}}(\ell \circ \mathsf{soft} \circ \mathcal{F}_{\beta,\nu}) \leq 2C \left(\phi_{\ell} \cdot (N_C - 1) \cdot R_s \cdot \xi(\boldsymbol{Y}) \cdot \sqrt{\frac{N_{par}\left(\nu N_L + \beta + \log\left(3\beta R_{\mathcal{X}} N\right)\right)}{N}} \right). \tag{B.54}$$

通过选择$\tilde{C} = 2C$即可完成证明。 □

附录 C　第5章中的证明

C.1　定理 5.3的证明

证明.

记 $\alpha = \frac{\alpha_3}{2C}$，有：

1) $\check{\delta}(\mathcal{L}_{\mathcal{G}_{BI}}) > \alpha > 0$.

根据(Boyd 等, 2004)的第5.9.2节，令约束$-\boldsymbol{U} \preceq 0$和$\boldsymbol{U} - \boldsymbol{I} \preceq 0$的拉普拉斯乘子分别为$\Omega_1$，$\Omega_2 \in \mathbb{R}^{(d+T)\times(d+T)}$。记 $tr(\boldsymbol{U}) = k$的拉普拉斯乘子为β。可得，KTT条件[1]为:

$$\mathcal{L}_{\mathcal{G}_{BI}} + \alpha\boldsymbol{U} - \Omega_1 + \Omega_2 + \beta\boldsymbol{I} = 0 \tag{C.1}$$

$$\langle -\boldsymbol{U}, \Omega_1 \rangle = 0, \langle \boldsymbol{U} - \boldsymbol{I}, \Omega_2 \rangle = 0, \tag{C.2}$$

$$\Omega_1 \succeq 0, \Omega_2 \succeq 0, \boldsymbol{U} \succeq 0, \boldsymbol{I} - \boldsymbol{U} \succeq 0. \tag{C.3}$$

记 Ω_1，Ω_2 和 $\boldsymbol{U}$的特征值分别为$\boldsymbol{\omega}_1 = diag(\omega_{1i})$, $\boldsymbol{\omega}_2 = diag(\omega_{2i})$, $\boldsymbol{\lambda} = diag(\lambda_i)$。由(Andreani 等, 2020)的引理2可得：

$$-\mathcal{L}_{\mathcal{G}_{BI}} = \boldsymbol{V}_U(\alpha\boldsymbol{\lambda} + \boldsymbol{\omega}_2 - \boldsymbol{\omega}_1 + \beta\boldsymbol{I})\boldsymbol{V}_U^\top, \tag{C.4}$$

$$\omega_{1i} \cdot \lambda_i = 0, \forall i \in [N], \tag{C.5}$$

$$\omega_{2i} \cdot (\lambda_i - 1) = 0, \forall i \in [N], \tag{C.6}$$

$$\gamma_i \geq 0, \omega_{1i} \geq 0, \omega_{2i} \geq 0, \ \ \forall\, i \in [N], \tag{C.7}$$

$$1 \geq \lambda_i \geq 0, \forall\, i \in [N], \tag{C.8}$$

$$tr(\boldsymbol{U}) = k. \tag{C.9}$$

其中 $\boldsymbol{V}_U$包含$\boldsymbol{U}$的特征向量。分以下几种情形对其分别进行证明：

情形(1)： $\boldsymbol{p = k - 1, q = k}$。说明目标谱间隙是非空的，则原始变量和对偶变量

[1]由于没有该约束是最优解是可行的，此处省略了约束$\boldsymbol{U} = \boldsymbol{U}^\top$

的解为:

$$\omega_{1i} = \left(\lambda_i\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \lambda_k\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \alpha\right) \cdot \mathbb{1}\left[i > k\right], \tag{C.10}$$

$$\omega_{2i} = \left(\lambda_k\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \lambda_i\left(\mathcal{L}_{\mathcal{G}_{BI}}\right)\right) \cdot \mathbb{1}\left[i \leq k\right], \tag{C.11}$$

$$\beta_i = -\lambda_k\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \alpha, \tag{C.12}$$

$$\boldsymbol{U}^\star = \sum_{i=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{C.13}$$

情形(2). 接下来证明，当$\boldsymbol{p \neq k-1, q \neq k}$时结论成立。具体而言，通过以下四种情形进行说明：

情形 (2a): $\boldsymbol{p \neq 0, q \neq N}$. 下列原始变量和对偶变量满足KKT条件：

$$\omega_{1i} = \left(\lambda_i\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \lambda_{p+1}\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \frac{k-p}{q-p}\alpha\right) \cdot \mathbb{1}\left[i > q\right], \tag{C.14}$$

$$\omega_{2i} = \left(\lambda_{p+1}\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \lambda_i\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \frac{q-k}{q-p}\alpha\right) \cdot \mathbb{1}\left[i \leq p\right], \tag{C.15}$$

$$\beta_i = -\lambda_{p+1}\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \frac{k-p}{q-p}\alpha, \tag{C.16}$$

$$\boldsymbol{U}^\star = \sum_{i=1}^{p} \boldsymbol{v}_i \boldsymbol{v}_i^\top + \frac{k-p}{q-p} \sum_{j=p+1}^{q} \boldsymbol{v}_j \boldsymbol{v}_j^\top. \tag{C.17}$$

情形 (2b): $\boldsymbol{p = 0, q \neq N}$。由图拉普拉斯矩阵的谱性质可知：

$$0 = \lambda_1(\mathcal{L}_{\mathcal{G}_{BI}}) = \cdots = \lambda_k(\mathcal{L}_{\mathcal{G}_{BI}}) = \cdots = \lambda_q(\mathcal{L}_{\mathcal{G}_{BI}}) \leq \lambda_{q+1}(\mathcal{L}_{\mathcal{G}_{BI}}) \leq \lambda_N(\mathcal{L}_{\mathcal{G}_{BI}}). \tag{C.18}$$

下列原始变量和对偶变量满足KKT条件:

$$\omega_{1i} = \left(\lambda_i\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \frac{k}{q}\alpha\right) \cdot \mathbb{1}\left[i > q\right], \tag{C.19}$$

$$\omega_{2i} = 0, \tag{C.20}$$

$$\beta_i = -\frac{k}{q}\alpha, \tag{C.21}$$

$$\boldsymbol{U}^\star = \frac{k}{q} \sum_{j=1}^{q} \boldsymbol{v}_j \boldsymbol{v}_j^\top. \tag{C.22}$$

情形 (2c): $\boldsymbol{p \neq 0, q = N}$。由$\mathcal{L}_{\mathcal{G}_{BI}} \neq \mathbf{0}$可知：

$$\lambda_1(\mathcal{L}_{\mathcal{G}_{BI}}) \leq \cdots \leq \lambda_p(\mathcal{L}_{\mathcal{G}_{BI}}) < \lambda_{p+1}(\mathcal{L}_{\mathcal{G}_{BI}}) = \cdots = \lambda_k(\mathcal{L}_{\mathcal{G}_{BI}}) = \cdots = \lambda_N(\mathcal{L}_{\mathcal{G}_{BI}}). \tag{C.23}$$

下列原始变量和对偶变量满足KKT条件:

$$\omega_{1i} = 0, \tag{C.24}$$

$$\omega_{2i} = \left(\lambda_{p+1}\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \lambda_i\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \frac{N-k}{N-p}\alpha\right) \cdot \mathbb{1}\left[i \le p\right], \tag{C.25}$$

$$\beta_i = -\lambda_{p+1}\left(\mathcal{L}_{\mathcal{G}_{BI}}\right) - \frac{k-p}{N-p}\alpha, \tag{C.26}$$

$$\boldsymbol{U}^\star = \sum_{i=1}^{p} \boldsymbol{v}_i \boldsymbol{v}_i^\top + \frac{k-p}{N-p} \sum_{j=p+1}^{N} \boldsymbol{v}_j \boldsymbol{v}_j^\top. \tag{C.27}$$

情形(2d): $\boldsymbol{p = 0, q = N}$。由于$\mathcal{L}_{\mathcal{G}_{BI}} \neq \mathbf{0}$，该情形不可能发生。

至此，说明了在所有$\alpha > 0$的情形下，有$\boldsymbol{U}^\star = \boldsymbol{V}\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^\top$。至此，对于$\alpha > 0$的情形，证明完毕。

2)$\boldsymbol{\alpha = 0}$。由于$\boldsymbol{U}^\star$是一个可行解，足以说明Eq.(5.11)是原问题的一个最优解。根据定理5.2有：

$$\boldsymbol{U}^\star \in \underset{\boldsymbol{U}\in\Gamma}{\operatorname{argmin}} \left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U} \right\rangle \quad 若 \quad \left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U}^\star \right\rangle = \sum_{i=1}^{k} \lambda_i(\mathcal{L}_{\mathcal{G}_{BI}}). \tag{C.28}$$

进一步，有：

$$\begin{aligned}
\left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U}^\star \right\rangle &= tr\left(\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^\top \mathcal{L}_{\mathcal{G}_{BI}} \boldsymbol{V}\right) \\
&= tr\left(\tilde{\boldsymbol{\Lambda}}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top \boldsymbol{V}\right) \\
&= tr\left(\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Lambda}\right) = \sum_{i=1}^{p} \lambda_i(\mathcal{L}_{\mathcal{G}_{BI}}) + \frac{k-p}{q-p} \sum_{i=p+1}^{q} \lambda_{p+1}(\mathcal{L}_{\mathcal{G}_{BI}}) \\
&= \sum_{i=1}^{p} \lambda_i(\mathcal{L}_{\mathcal{G}_{BI}}) + (k-p) \cdot \lambda_{p+1}(\mathcal{L}_{\mathcal{G}_{BI}}) = \sum_{i=1}^{k} \lambda_i(\mathcal{L}_{\mathcal{G}_{BI}}).
\end{aligned} \tag{C.29}$$

至此，**2)**证毕。 □

C.2　非凸-非光滑优化的预备知识

C.2.1　次梯度

现在介绍本征下半连续函数(不必为凸函数)的广义次梯度(Rockafellar 等, 2009)，这是后续收敛性分析的基础。

定义 C.1. 考虑函数$f : \mathbb{R}^n \to \bar{\mathbb{R}}$和点$\bar{x}$，其中$f(\bar{x})$是有限的。对于任意向量$v \in \mathbb{R}^n$，有：

(1) 若

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x} \rangle + o(|x - \bar{x}|); \tag{C.30}$$

则v是f在$\bar{x}$处的**常规次梯度**，记作$v \in \hat{\partial} f(\bar{x})$。

(2) 若存在序列$x^\nu \to \bar{x}$，满足$f(x^\nu) \to f(\bar{x})$，且存在v使得$v^\nu \in \hat{\partial} f(x^\nu)$满足$v^\nu \to v$，那么$v$是$f$在$\bar{x}$处的**广义次梯度**，记作$v \in \partial f(\bar{x})$。

注意a)中的符号o是单向极限条件的简写:

$$\liminf_{x \to \bar{x}, x \neq \bar{x}} \frac{f(x) - f(\bar{x}) - \langle v, x - \bar{x} \rangle}{|x - \bar{x}|} \geq 0 \tag{C.31}$$

具体而言，广义次梯度具有下列性质：

性质 C.1 (次梯度的存在性). ((Rockafellar 等, 2009)的推论8.10) 若函数$f : \mathbb{R}^n \to \bar{\mathbb{R}}$正定且在$\bar{x}$处下半连续，则$\partial f(\bar{x})$非空。

基于性质1可知，对于正定且下半连续的函数，广义次梯度总是存在。

性质 C.2 (局部极小值的广义费马规则). (定理10.1 of (Rockafellar 等, 2009)) 若本征函数$f : \mathbb{R}^n \to \bar{\mathbb{R}}$在$\bar{x}$处有一个局部极小值，则 $0 \in \partial f(\bar{x})$。

根据性质2，当$0 \in \partial f$时，定义$\bar{x}$为f的一个**临界点**。

性质 C.3 (广义次梯度). ((Rockafellar 等, 2009)的命题 8.12) 若本征函数$f : \mathbb{R}^n \to \bar{\mathbb{R}}$是凸的，则有：

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{v} : f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle, \boldsymbol{y} \in \mathbb{R}^n\} = \hat{\partial} f(\boldsymbol{x}). \tag{C.32}$$

以上性质表明，次梯度的定义与凸函数的定义相兼容。

C.2.2 KL 函数

定义 C.2 (Kurdyka-Łojasiewicz (KL) 性质(Attouch 等, 2010; Zeng 等, 2019)). 对于任意函数$G : \mathbb{R}^n \to \mathbb{R} \cup +\infty$，其自变量$\boldsymbol{x} \in dom\{\partial G\}$。若存在$\eta \in (0, +\infty)$，存在一个以 $\boldsymbol{x}$ 为球心，ρ为半径的开球$\mathcal{B}(\boldsymbol{x}, \rho)$，且存在一个凹函数 $\phi(t)$ 满足：(1) 在0处连续，(2) $\phi(0) = 0$，(3) $\phi \in \mathcal{C}^1((0, \eta))$，(4) $\phi'(x) > 0, \forall\, x \in (0, \eta)$，从而对于任意$\boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}, \rho) \cap [G(\boldsymbol{x}) < G(\boldsymbol{y}) < G(\boldsymbol{x}) + \eta]$, 下述 KL不等式成立:

$$\phi'\big(G(\boldsymbol{y}) - G(\boldsymbol{x})\big) \cdot dist\big(0, \partial G(\boldsymbol{y})\big) \geq 1, \tag{C.33}$$

其中，对于集合$\mathcal{S} \subset \mathcal{R}^n$, $dist(\boldsymbol{x}, \mathcal{S}) = \inf_{\boldsymbol{y} \in \mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|$， **则称该函数$G$在$\boldsymbol{x} \in dom\{\partial G\}$处具有KL性质**。

定义 C.3 (KL 函数(Attouch 等, 2010; Zeng 等, 2019)). 在$dom\{\partial G\}$的每一个点处满足Kurdyka-Łojasiewicz 不等式的本征下半连续函数被称为KL 函数。

本章使用了两类 KL函数: 半代数函数和可定义函数。首先给出半代数函数和半代数集的定义。

定义 C.4 (半代数函数(Bochnak 等, 2013; Zeng 等, 2019; Fu 等, 2019)). 半代数集合和半代数函数的定义分别为:

(1) 若一个集合$\mathcal{A} \subset \mathbb{R}^n$能表示为:

$$\mathcal{A} = \bigcup_{i=1}^{m} \bigcap_{j=1}^{n} \left\{ \boldsymbol{x} \in \mathbb{R}^n : p_{ij}(\boldsymbol{x}) = 0, q_{ij}(\boldsymbol{x}) > 0 \right\},$$

其中 p_{ij}, q_{ij}是实多项式函数，且$i \in [m], j \in [n]$，则称该集合为半代数集。

(2) 若一个函数的象

$$Gr(h) = \{(x, h(x)) : x \in dom(h)\}$$

是一个半代数集，则称其为半代数函数。

在介绍可定义函数之前，首先需要了解半代数的一个扩展概念，即o-最小结构(Bolte 等, 2007)，其定义如下:

定义 C.5 (o-最小结构). 一个$(\mathbb{R}, +, \cdot)$上的o-最小结构$\mathcal{O}_n$是$\mathbb{R}^n$的可定义子集的一个布尔代数序列，从而对于每个 $n \in \mathbb{N}$，有:

(1) 若A属于$\mathcal{O}_n$，那么，$A \times \mathbb{R}$和$\mathbb{R} \times A$均属于 $\mathcal{O}_{n+1}$;

(2) 若$\Pi : \mathbb{R}_{n+1} \to \mathbb{R}_n$是到$\mathbb{R}_n$的正则投影，则对于任意$A$属于$\mathcal{O}_{n+1}$，集合$\Pi(A)$属于$\mathcal{O}_n$;

(3) $\mathcal{O}_n$包含$\mathbb{R}_n$的代数子集，即，每个集合均形如:

$$\{x \in \mathbb{R}_n : p(x) = 0\},$$

其中 $p : \mathbb{R}_n \to \mathbb{R}$ 是一个多项式函数。

(4) $\mathcal{O}_1$的元素是区间和点的有限并集。

基于o-最小结构的定义，可以方便给出可定义函数的定义。

定义 C.6 (可定义函数). 给定一个$((\mathbb{R}, +, \cdot)$上的$)o$-最小结构 $\mathcal{O}$, 函数 $f : \mathbb{R}_n \to \mathbb{R}$被称为是$\mathcal{O}$ 上可定义的，当其象属于$\mathcal{O}_{n+1}$。

注 C.1. 根据(Van den Dries 等, 1996; Bolte 等, 2007)，下面给出关于o-最小结构的一些重要事实。

(1) 半代数集列是一个o-最小结构。回顾前文，半代数集是集合的布尔组合，形如：

$$\{x \in \mathbb{R}_n : p(x) = 0, q_1(x) < 0 \dots q_m(x) < 0\},$$

其中 p和q_i是$\mathbb{R}_n$上的多项式函数。

(2) 在求和、复合、反卷积和其他几种经典分析操作下，o-最小结构是稳定的。

下列关于半代数集、半代数函数和可定义函数的性质对后续的分析至关重要：

命题 C.1. 以下事实成立:

(1) 半代数集的指示函数是半代数函数(Lau 等, 2018)。

(2) 半代数函数之间的有限和与有限积是半代数函数(Lau 等, 2018)。

(3) 半代数集的交集与有限并集是半代数集(Lau 等, 2018)。

(4) 多项式函数是半代数函数。

(5) 半代数函数是可定义函数。

(6) 可定义函数是KL 函数(Bolte 等, 2007)。

证明. **证明(4)** 对于任意多项式函数 $y = h(x) = p_n(x)$，其象可被重新形式化为：

$$Gr(h) = \{(x, y) : y - p_n(x) = 0, y - p_n(x) + 1 > 0\}. \tag{C.34}$$

显然，$y - p_n(x)$和$y - p_n(x) + 1$ 都是实多项式，证毕。

证明(5) 遵循 Rem. C.1-(1)。 □

C.3 证明TFCL优化算法的收敛性

基于上节的预备知识，本小节证明定理5.4和定理5.5。在本节，记替代问题$(\boldsymbol{P}^\star)$的总体目标函数为$\mathcal{F}(\boldsymbol{W}, \boldsymbol{U})$，记原问题 $(\boldsymbol{P})$的对应目标函数为 $\widetilde{\mathcal{F}}(\boldsymbol{W}, \boldsymbol{U})$。

首先分析 $(\boldsymbol{P}^\star)$问题。

引理 C.1. 令 $\mathcal{F}(\boldsymbol{W}, \boldsymbol{U}) = \mathcal{J}(\boldsymbol{W}) + \alpha \cdot \left\langle \mathcal{L}_{\mathcal{G}_{BI}}, \boldsymbol{U} \right\rangle + \iota_\Gamma(\boldsymbol{U})$，其中 $\iota_\Gamma(\cdot)$ 是集合Γ的指示函数，令 $\boldsymbol{W}^t, \boldsymbol{U}^t$表示第$t$轮学得的参数。对于算法 6，若$\mathcal{J}(\boldsymbol{W})$是一个可定义函数, $\nabla_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W})$是 ϱ-Lipschitz 连续的，且$\boldsymbol{W}_t \neq 0, \forall t$，则有下述性质成立：

(1) (充分下降条件): 如果 $C > \varrho$，序列$\{\mathcal{F}(\boldsymbol{W}^t, \boldsymbol{U}^t)\}$ 是非递增的：

$$\mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1}) \leq \mathcal{F}(\boldsymbol{W}^t, \boldsymbol{U}^t) - \min\left\{\frac{C-\varrho}{2}, \frac{\alpha_3}{2}\right\} \|\Delta(\boldsymbol{\Theta}^t)\|^2.$$

其中 $\Delta(\boldsymbol{\Theta}^t) = \left[vec(\Delta(\boldsymbol{W})^t); vec(\Delta(\boldsymbol{U}^t))\right]$，$\Delta(\boldsymbol{W}^t) = \boldsymbol{W}^{t+1} - \boldsymbol{W}^t$，$\Delta(\boldsymbol{U}^t) = \boldsymbol{U}^{t+1} - \boldsymbol{U}^t$。

(2) (平方可加性): $\sum_{i=1}^{\infty} ||\Delta(\boldsymbol{\Theta}^t)||_F^2 < \infty$。进一步，有 $\lim_{t\to\infty} ||\Delta(\boldsymbol{U}^t)|| = 0$，以及$\lim_{t\to\infty} ||\Delta(\boldsymbol{W}^t)|| = 0$。

(3) (连续条件):存在一个$\{\boldsymbol{W}^{k_j}, \boldsymbol{U}^{k_j}\}_j$ 的子序列，存在一个聚点$\{\boldsymbol{W}^\star, \boldsymbol{U}^\star\}$，满足：

$$\{\boldsymbol{W}^{k_j}, \boldsymbol{U}^{k_j}\} \to \{\boldsymbol{W}^\star, \boldsymbol{U}^\star\}, \mathcal{F}(\boldsymbol{W}^{k_j}, \boldsymbol{U}^{k_j}) \to \mathcal{F}(\boldsymbol{W}^\star, \boldsymbol{U}^\star).$$

(4) (KL 性质)：$\mathcal{F}(\cdot,\cdot)$ 是一个 KL 函数。

(5) (相对误差条件)：对于任意$t \in \mathbb{N}$，下式成立：

$$dist(\mathbf{0}, \partial_\Theta \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})) \leq \left[C + \varrho + \alpha_1(\sqrt{d} + \sqrt{T} + 2)\right] \cdot \left\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t\right\|_F. \quad \text{(C.35)}$$

C.3.1　引理C.1的证明

证明.

证明(1):

由于$\mathcal{J}(\cdot)$是ϱ-Lipschitz 连续函数，对于第$t+1$轮迭代:

$$\mathcal{J}(\boldsymbol{W}^{t+1}) \leq \mathcal{J}(\boldsymbol{W}^t) + \left\langle \nabla_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W}^t), \Delta(\boldsymbol{W}^t)\right\rangle + \frac{\varrho}{2}\|\Delta(\boldsymbol{W}^t)\|_F^2. \quad \text{(C.36)}$$

将$\boldsymbol{W}^t$带入$\boldsymbol{W}$， $\boldsymbol{U}$ 子问题的唯一解为$\boldsymbol{U}^{t+1}$，由于子问题是 α_3-强凸的，有：

$$\begin{aligned} &\alpha_1 \cdot \left\langle \boldsymbol{D}^{t+1}, |\boldsymbol{W}^t|\right\rangle + \iota_\Gamma(\boldsymbol{U}^{t+1}) + \frac{\alpha_3}{2}\|\boldsymbol{U}^{t+1}\|_F^2 \\ \leq &\alpha_1 \cdot \left\langle \boldsymbol{D}^t, |\boldsymbol{W}^t|\right\rangle + \iota_\Gamma(\boldsymbol{U}^t) + \frac{\alpha_3}{2}\|\boldsymbol{U}^t\|_F^2 - \frac{\alpha_3}{2}||\Delta\boldsymbol{U}^t||_F^2. \end{aligned} \quad \text{(C.37)}$$

对于$\boldsymbol{W}$ 子问题，有:

$$\underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{W} - \widetilde{\boldsymbol{W}^t}\right\|_F^2 + \frac{\alpha_1}{C} \cdot \left\langle \boldsymbol{D}^{t+1}, |\boldsymbol{W}|\right\rangle + \frac{\alpha_2}{2C}\|\boldsymbol{W}\|_F^2, \quad \text{(C.38)}$$

可知该问题是强凸的，说明解$\boldsymbol{W}^{k+1}$ 是该子问题的最小值。从而：

$$\begin{aligned}&\left\langle \nabla_{\boldsymbol{W}} \mathcal{J}(\boldsymbol{W}^t), \Delta \boldsymbol{W}^t \right\rangle + \frac{C}{2}\|\Delta \boldsymbol{W}^t\|_F^2 + \alpha_1 \cdot \left\langle \boldsymbol{D}^{t+1}, |\boldsymbol{W}^{t+1}| \right\rangle + \frac{\alpha_2}{2}\|\boldsymbol{W}^{t+1}\|_F^2 \\ \leq &\alpha_1 \cdot \left\langle \boldsymbol{D}^{t+1}, |\boldsymbol{W}^t| \right\rangle + \frac{\alpha_2}{2}\|\boldsymbol{W}^t\|_F^2.\end{aligned} \tag{C.39}$$

综合(C.36)， (C.37)，(C.39)可得：

$$\mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1}) \leq \mathcal{F}(\boldsymbol{W}^t, \boldsymbol{U}^t) - \min\{\frac{C-\varrho}{2}, \frac{\alpha_3}{2}\} \cdot ||\Delta(\Theta^t)||_F^2, \tag{C.40}$$

至此完成了1)的证明。

证明(2):

将(C.40)， $t = 1, 2, \cdots$ 进行求和，基于事实$\mathcal{F}(\cdot, \cdot) \geq 0$，有

$$\sum_{t=1}^{\infty} \min\{\frac{C-\varrho}{2}, \frac{\alpha_3}{2}\} \cdot ||\Delta(\boldsymbol{\Theta}^t)||_F^2 \leq \mathcal{F}(\boldsymbol{W}^0, \boldsymbol{U}^0). \tag{C.41}$$

根据假设$\mathcal{F}(\boldsymbol{W}^0, \boldsymbol{U}^0) < \infty$，从而直接证明：$\Delta(\boldsymbol{W}^k) \overset{t\to\infty}{\to} 0$，且$\Delta(\boldsymbol{U}^t) \overset{t\to\infty}{\to} 0$。

证明(3):

接下来证明序列$\{\boldsymbol{W}^t, \boldsymbol{U}^t\}_t$存在聚点。对所有$t \in \mathbb{N}$，由于$\boldsymbol{U}_t \in \Gamma$, 故$\{\boldsymbol{U}^t\}$是有界的。类似地，由于损失序列$\{\mathcal{F}(\boldsymbol{W}^t, \boldsymbol{U}^t)\}$非递增，有$\|\boldsymbol{W}^t\| \leq \sqrt{\frac{2\mathcal{F}(\boldsymbol{U}^0, \boldsymbol{W}^0)}{\alpha_2}}$，故序列$\{\boldsymbol{W}^t\}$是有界的。根据Bolzano-Weierstrass定理，任意一个有界序列必定有收敛子列，从而直接证明$\{\boldsymbol{W}^t, \boldsymbol{U}^t\}$中至少存在一个聚点。

下面证明连续条件。选取任意一个收敛子序列$\{\boldsymbol{W}^{t_j}, \boldsymbol{U}^{t_j}\}_j$，设其聚点为$\boldsymbol{W}^\star, \boldsymbol{U}^\star$。假设$\mathcal{J}$是连续函数，可知$\mathcal{F}(\boldsymbol{W}, \boldsymbol{U}) - \iota(\boldsymbol{U})$也是连续的，必定有

$$\lim_{j\to} \mathcal{F}(\boldsymbol{W}^{k_j}, \boldsymbol{U}^{k_j}) - \iota(\boldsymbol{U}^{k_j}) = \mathcal{F}(\boldsymbol{W}^\star, \boldsymbol{U}^\star) - \iota(\boldsymbol{U}^\star)$$

从而只需证明$\lim_{j\to\infty} \iota(\boldsymbol{U}^{k_j}) = \iota(\boldsymbol{U}^\star)$。由于$\iota(\cdot)$是下半连续函数，可知

$$\liminf_{j\to\infty} \iota(\boldsymbol{U}^{k_j}) \geq \iota(\boldsymbol{U}^\star)$$

从而只需证$\limsup_{j\to\infty}\iota(\boldsymbol{U}^{k_j})\le\iota(\boldsymbol{U}^\star)$。给定任意 t_j，固定$\boldsymbol{W}^{t_j-1}$，可知$\boldsymbol{U}^{t_j}$是$\boldsymbol{U}$-子问题的最优值，因此得到:

$$\begin{aligned}&\alpha_1\cdot\left\langle\boldsymbol{D}^{t_j},|\boldsymbol{W}^{t_j-1}|\right\rangle+\iota_\Gamma(\boldsymbol{U}^{t_j})+\frac{\alpha_3}{2}\|\boldsymbol{U}^{t_j}\|_F^2\\\le&\alpha_1\cdot\left\langle\boldsymbol{D}^\star,|\boldsymbol{W}^{t_j-1}|\right\rangle+\iota_\Gamma(\boldsymbol{U}^\star)+\frac{\alpha_3}{2}\|\boldsymbol{U}^\star\|_F^2.\end{aligned}\tag{C.42}$$

将上述不等式两边同时对j取极限$j\to\infty$，得:

$$\iota_\Gamma(\boldsymbol{U}^{t_j})\le\iota_\Gamma(\boldsymbol{U}^\star),\quad j\to\infty.\tag{C.43}$$

从而得到$\limsup_{j\to\infty}\iota(\boldsymbol{U}^{k_j})\le\iota(\boldsymbol{U}^\star)$。

证明(4):

根据命题 C.1，多项式函数是半代数函数，从而可知正则化$\frac{\alpha_2}{2}\|\boldsymbol{W}\|_F^2$，$\frac{\alpha_3}{2}\|\boldsymbol{U}\|_F^2$s是半代数函数。

现在证明$\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle$ 是半代数函数。由于

$$\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle=\sum_{i=1}^{T}\sum_{j=d}^{d+T}\mathcal{L}_{\mathcal{G}_{BI}ij}U_{ij}+\sum_{i=d}^{d+T}\sum_{j=1}^{d}\mathcal{L}_{\mathcal{G}_{BI}ij}U_{ij},$$

且

$$\mathcal{L}_{\mathcal{G}_{BI}ij}U_{ij}=\begin{cases}\left[\mathbb{1}[i=j]\cdot\left(\sum_{k=1}^{T}|W|_{ik}U_{ij}\right)\right]-|W|_{i,j-d}U_{ij} & ,i\le d,j>d,\\\left[\mathbb{1}[i=j]\cdot\left(\sum_{k=1}^{d}|W|_{kj}U_{ij}\right)\right]-|W|_{j,i-d}U_{ij} & ,i>d,j\le d,,\\0 & ,otherwise,\end{cases}$$

因为$\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle$ 形如$y=|x_1|\cdot x_2$的函数的求和，综合命题 C.1可知，当$y=|x_1|\cdot x_2$是半代数函数时，$\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle$也是半代数函数。此函数的象可形式化为

$$\{(x_1,x_2,y):y+x_1\cdot x_2=0,x_1<0\}\cup\{(x_1,x_2,y):y-x_1\cdot x_2=0,x_1>0\},$$

从而易知其是半代数函数，故 $\left\langle\mathcal{L}_{\mathcal{G}_{BI}},\boldsymbol{U}\right\rangle$ 是半代数函数。

现在说明指示函数 $\iota_\Gamma(\boldsymbol{U})$是半代数函数。根据命题 C.1，只需说明$\Gamma$ 是半代数集。易将Γ重新形式化为：

$$\Gamma=\{\boldsymbol{U}:\boldsymbol{U}=\boldsymbol{U}^\top,\ \boldsymbol{U}\succeq\boldsymbol{0},\ \boldsymbol{I}-\boldsymbol{U}\succeq\boldsymbol{0},\ tr(\boldsymbol{U})=k\}.\tag{C.44}$$

显然，

$$\begin{aligned}\Gamma_1 &= \{\boldsymbol{U} : \boldsymbol{U} = \boldsymbol{U}^\top\} \\ &= \cap_{1\le i\neq j\le n}\{U_{ij} : U_{ij} = U_{ji}\} = \cap_{1\le i\neq j\le n}\{U_{ij} : U_{ij} = U_{ji},\, U_{ij} - U_{ji} + 1 > 0\}.\end{aligned} \tag{C.45}$$

故Γ_1 是半代数集。

对于Γ_2，有

$$\Gamma_2 = \{\boldsymbol{U} : tr(\boldsymbol{U}) = k\} = \{\boldsymbol{U} : tr(\boldsymbol{U}) = k, tr(\boldsymbol{U}) - k + 1 > 0\}, \tag{C.46}$$

证明Γ_2 是半代数集。

现在证明$\Gamma_3 = \{\boldsymbol{U} \in \mathbb{R}^{N\times N} : \boldsymbol{U} \succeq 0\}$是半代数集。根据半正定矩阵的基本性质可知$\boldsymbol{U} \succeq 0$当且仅当所有主子式都是非负的。具体来说，给定矩阵$\boldsymbol{U}$，由矩阵$\boldsymbol{U}$第$k_1, k_2, \cdots k_l$行上的所有第$k_1, k_2, \cdots k_l$个元素组成的$l \times l$维矩阵$\boldsymbol{U}_{k_1,k_2,\cdots k_l}$，称为主子矩阵($1 \le l \le N$和$k_1, k_2, \cdots k_l \in \{1, 2, \cdots l\}$)。此外，$\boldsymbol{U}_{k_1,k_2,\cdots k_l}$的行列式被称为$\boldsymbol{U}$的主子式。从而，对任意$l$，任意$k_1, k_2, \cdots k_l$，$\boldsymbol{U} \succeq 0$等价于$-Det(\boldsymbol{U})_{k_1,k_2,\cdots k_l} \le 0$，可被重写为:

$$\boldsymbol{U} \in \bigcap_{1\le l\le N}\ \bigcap_{k_1,k_2,\cdots k_l} \{\boldsymbol{U} : -Det(\boldsymbol{U}_{k_1,k_2,\cdots k_l}) \le 0\}. \tag{C.47}$$

根据行列式的定义，$Det(\boldsymbol{U}_l)$可被表示为一个以$\boldsymbol{U}_l$为元素的多项式函数，从而Γ_3是半代数集(0是实数多项式函数)。$\Gamma_4 = \{\boldsymbol{U} \in \mathbb{R}^{N\times N} : \boldsymbol{I} - \boldsymbol{U} \succeq \boldsymbol{0}\}$ 的证明同Γ_3的证明。

根据命题 C.1, $\Gamma = \Gamma_1 \cap \Gamma_2 \cap \Gamma_3 \cap \Gamma_4$是半代数集，$\iota_\Gamma(\cdot)$同理。

由于半代数函数是可定义函数，$\mathcal{F}$的所有求和是可定义函数，证明了$\mathcal{F}(\boldsymbol{W}, \boldsymbol{U})$ 是一个KL 函数。

证明(5):

根据$\boldsymbol{U}^{t+1}$关于$\boldsymbol{U}$子问题在第 $t+1$轮迭代的最优性，可知：

$$\begin{aligned}\boldsymbol{0} &\in \alpha_1 \cdot \mathcal{L}^t_{\mathcal{G}_{BI}} + \partial_{\boldsymbol{U}} \iota_\Gamma(\boldsymbol{U}^{t+1}) + \alpha_3 \boldsymbol{U}^{t+1} \\ &= \partial_{\boldsymbol{U}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1}) + \alpha_1 \cdot (\mathcal{L}^{t+1}_{\mathcal{G}_{BI}} - \mathcal{L}^t_{\mathcal{G}_{BI}}).\end{aligned} \tag{C.48}$$

相应地，有 $\exists \boldsymbol{g}_{\boldsymbol{U}} \in \partial_{\boldsymbol{U}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})$，从而

$$\|\boldsymbol{g}_{\boldsymbol{U}}\| = \alpha_1 \cdot ||\mathcal{L}^{t+1}_{\mathcal{G}_{BI}} - \mathcal{L}^{t}_{\mathcal{G}_{BI}}|| \tag{C.49}$$

$$\leq \alpha_1 \cdot \left(\left\| diag(|\boldsymbol{W}^t|\mathbf{1}) - diag(|\boldsymbol{W}^{t+1}|\mathbf{1}) \right\| + \left\| diag(|\boldsymbol{W}^{t\top}|\mathbf{1}) - diag(|\boldsymbol{W}^{t+1\top}|\mathbf{1}) \right\| \right) \tag{C.50}$$

$$+ \alpha_1 \cdot \left(2 \cdot \left\| |\boldsymbol{W}^t| - |\boldsymbol{W}^{t+1}| \right\| \right) \tag{C.51}$$

$$\leq \alpha_1 (\sqrt{k} + \sqrt{T} + 2) \cdot \|\boldsymbol{W}^t - \boldsymbol{W}^{t+1}\|. \tag{C.52}$$

因此，

$$dist(\mathbf{0}, \partial_{\boldsymbol{U}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})) \leq \alpha_1 (\sqrt{d} + \sqrt{T} + 2) \cdot \|\boldsymbol{W}^t - \boldsymbol{W}^{t+1}\|. \tag{C.53}$$

根据$\boldsymbol{W}^{t+1}$关于$\boldsymbol{W}$ 子问题在第$t+1$轮迭代的最优性可知：

$$\begin{aligned} \mathbf{0} &\in \nabla_W \mathcal{J}(\boldsymbol{W}^t) + C \cdot (\boldsymbol{W}^{t+1} - \boldsymbol{W}^t) + \\ &\alpha \cdot \partial_{\boldsymbol{W}} \boldsymbol{\Omega}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1}) + \alpha_2 \cdot \boldsymbol{W}^{t+1} \\ &= \partial_{\boldsymbol{W}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1}) \\ &+ C \cdot (\boldsymbol{W}^{t+1} - \boldsymbol{W}^t) + \nabla_W \mathcal{J}(\boldsymbol{W}^t) - \nabla_W \mathcal{J}(\boldsymbol{W}^{t+1}). \end{aligned} \tag{C.54}$$

相应地， $\exists \boldsymbol{g}_{\boldsymbol{W}} \in \partial_{\boldsymbol{W}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})$，从而

$$||\boldsymbol{g}_{\boldsymbol{W}}|| = ||C \cdot (\boldsymbol{W}^{t+1} - \boldsymbol{W}^t) + \nabla_W \mathcal{J}(\boldsymbol{W}^t) - \nabla_W \mathcal{J}(\boldsymbol{W}^{t+1})||.$$

由于$\mathcal{J}$ 是ϱ-Lipschitz 连续的，有：

$$||\boldsymbol{g}_{\boldsymbol{W}}|| \leq (C + \varrho) \cdot ||\boldsymbol{W}^{t+1} - \boldsymbol{W}^t||.$$

因此

$$dist(\mathbf{0}, \partial_W \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})) \leq (C + \varrho) \cdot ||\boldsymbol{W}^{t+1} - \boldsymbol{W}^t||. \tag{C.55}$$

根据公式(C.53)和公式(C.55)有：

$$\begin{aligned} &dist(\mathbf{0}, \partial_\Theta(\mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})) \\ &\leq dist(\mathbf{0}, \partial_{\boldsymbol{U}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})) + dist(\mathbf{0}, \partial_{\boldsymbol{W}} \mathcal{F}(\boldsymbol{W}^{t+1}, \boldsymbol{U}^{t+1})) \\ &\leq \left[C + \varrho + \alpha_1 (\sqrt{d} + \sqrt{T} + 2) \right] \cdot \left\| \boldsymbol{W}^{t+1} - \boldsymbol{W}^t \right\|_F \\ &\leq \left[C + \varrho + \alpha_1 (\sqrt{d} + \sqrt{T} + 2) \right] \cdot \left\| \boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^t \right\|_F . \end{aligned}$$

□

C.3.2 定理5.4的证明

证明. 由(Attouch 等, 2013)的引理2.6，根据充分下降条件、连续条件、KL 性质和相对误差条件，(1)成立(见引理C.1)。

根据引理C.1-(1)，损失函数 $\mathcal{F}$ 是非递增的，下界为0，因此，$\{\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t)\}_t$收敛。进一步，由于$(\boldsymbol{W}^t,\boldsymbol{U}^t) \overset{t\to\infty}{\to} (\boldsymbol{W}^\star,\boldsymbol{U}^\star)$，$(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$是参数序列的唯一聚点，根据C.1-(3)，有 $\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t) \overset{t\to\infty}{\to} \mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$。从而2)得证。

根据引理C.1-(5)有：

$$\sum_{i=1}^{T} dist(\mathbf{0}, \partial_\Theta \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t))^2 \le +\infty, \tag{C.56}$$

从而证明(3)。 □

最后，基于定理5.2 和定理5.4，证明所提出算法关于原问题 $(\boldsymbol{P})$的收敛性。

C.3.3 定理5.5的证明

证明. **(1)的证明**

根据定理5.4可知$(\boldsymbol{W}^t,\boldsymbol{U}^t) \overset{t\to\infty}{\to} (\boldsymbol{W}^\star,\boldsymbol{U}^\star)$和$\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t) \overset{t\to\infty}{\to} \mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$。由此足以证明$(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$为$\widetilde{\mathcal{F}}$的临界点。

由引理C.1-(5)可知，$dist(\mathbf{0}, \partial_{\boldsymbol{W}} \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t)) \to 0$，$dist(\mathbf{0}, \partial_{\boldsymbol{U}} \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t)) \to 0$。由此可知：

$$0 \in \partial_{\boldsymbol{W}} \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t),\ t \to \infty, \tag{C.57}$$

且

$$0 \in \partial_{\boldsymbol{U}} \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t),\ \ t \to \infty. \tag{C.58}$$

$\boldsymbol{U}$固定时，$\mathcal{F}(\boldsymbol{W},\boldsymbol{U})$相对于$\boldsymbol{W}$是凸函数。同时，$\mathcal{F}(\boldsymbol{W},\boldsymbol{U})$在$\boldsymbol{W}$固定时相对于$\boldsymbol{U}$为凸函数。综合性质 C.3、式 (C.57)及式 (C.58)可得：

$$\lim_{t\to\infty} \mathcal{F}(\boldsymbol{W},\boldsymbol{U}^t) - \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t) \ge 0 \tag{C.59}$$

且

$$\lim_{t\to\infty} \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}) - \mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t) \ge 0 \tag{C.60}$$

由于$\mathcal{F}$相对于$\boldsymbol{W}$连续，故有 $\lim_{t\to\infty}\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U})=\mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U})$。此外，由于

$$\lim_{t\to\infty}\mathcal{F}(\boldsymbol{W},\boldsymbol{U}^t)-\iota_\Gamma(\boldsymbol{U}^t)=\mathcal{F}(\boldsymbol{W},\boldsymbol{U}^\star)-\iota_\Gamma(\boldsymbol{U}^\star)$$

且

$$\lim_{t\to\infty}\iota_\Gamma(\boldsymbol{U}^t)=\iota_\Gamma(\boldsymbol{U}^\star)$$

（已在引理C.1-(3)中证明$\boldsymbol{U}^\star$是唯一聚点）有 $\lim_{t\to\infty}\mathcal{F}(\boldsymbol{W},\boldsymbol{U}^t)=\mathcal{F}(\boldsymbol{W},\boldsymbol{U}^\star)$。结合已知条件$\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t)\overset{t\to\infty}{\to}\mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$，可得：

$$0\in\partial_{\boldsymbol{W}}\mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U}^\star),$$

且

$$0\in\partial_{\boldsymbol{U}}\mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U}^\star).$$

显然，$\partial_{\boldsymbol{W}}\mathcal{F}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)=\partial_{\boldsymbol{W}}\widetilde{\mathcal{F}}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$，即$0\in\partial_{\boldsymbol{W}}\widetilde{\mathcal{F}}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$。因为$0<\alpha_3<2C\min_t\check{\delta}(\mathcal{L}^t_{\mathcal{G}_{BI}})$，所以$\boldsymbol{U}^\star$是$U$-子问题的唯一最优解。进一步，$\boldsymbol{U}^\star$ 满足定理5.3，且是$\min_{\boldsymbol{U}}\widetilde{\mathcal{F}}(\boldsymbol{W}^\star,\boldsymbol{U})$的解。因此，$0\in\partial_{\boldsymbol{U}}\widetilde{\mathcal{F}}(\boldsymbol{W}^\star,\boldsymbol{U}^\star)$。综上所述，可证：

$$0\in\partial\widetilde{\mathcal{F}}(\boldsymbol{W}^\star,\boldsymbol{U}^\star).\tag{C.61}$$

(2)和(3)的证明

同(1)的证明逻辑相似，从

$$0\in\partial_{\boldsymbol{W}}\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t),\ \ 0\in\partial_{\boldsymbol{U}}\mathcal{F}(\boldsymbol{W}^t,\boldsymbol{U}^t)$$

可推出

$$0\in\partial_{\boldsymbol{W}}\widetilde{\mathcal{F}}(\boldsymbol{W}^t,\boldsymbol{U}^t),\ \text{和}\ \ 0\in\partial_{\boldsymbol{U}}\widetilde{\mathcal{F}}(\boldsymbol{W}^t,\boldsymbol{U}^t)\tag{C.62}$$

以下可通过定理5.4和引理C.1的方法证明(2)和(3)。 □

C.4　分组效应的证明

引理 C.2.　(a) 假设对于所有$\infty>\kappa>0$有$\sup_{\|\boldsymbol{W}\|_F\le\kappa}\|\nabla_{\boldsymbol{W}}\mathcal{J}(\boldsymbol{W})\|_\infty\le\varpi(\kappa)<\infty$成立。若算法6于第$\mathcal{T}$轮迭代停止，则：

$$\mathsf{Supp}(\boldsymbol{W}^{\mathcal{T}})\subseteq\left\{(i,j):\|\boldsymbol{f}_i^{\mathcal{T}}-\boldsymbol{f}_{d+j}^{\mathcal{T}}\|_2^2<\delta_1\right\}\tag{C.63}$$

其中

$$\delta_1=\frac{C}{\alpha_1}\cdot\left(C_0+\frac{\varpi(C_0)}{C}\right)=\frac{C}{\alpha_1}\kappa_0,\ C_0=\left(\frac{2}{\alpha_2}\cdot\epsilon_{\mathcal{T}-1}\right)^{1/2}.$$

(b) 若$\min_{(i,j)}|\widetilde{W}_{i,j}^{\mathcal{T}}| \geq \delta_0 > 0$，则：

$$\left\{(i,j) : ||\boldsymbol{f}_i^{\mathcal{T}} - \boldsymbol{f}_{d+j}^{\mathcal{T}}||_2^2 < \delta_2\right\} \subseteq \mathsf{Supp}(\boldsymbol{W}^{\mathcal{T}}) \tag{C.64}$$

其中 $\delta_2 = \dfrac{C}{\alpha_1} \cdot \delta_0$。

证明. **(a)的证明：**

由于$\widetilde{W}_{ij}^{\mathcal{T}} = W_{ij}^{\mathcal{T}-1} - \dfrac{[\nabla_{\boldsymbol{W}}\mathcal{J}(\boldsymbol{W}^{\mathcal{T}-1})]_{ij}}{C}$，根据假设$\sup_{\|\boldsymbol{W}\|_F \leq \kappa}\|\nabla_{\boldsymbol{W}}\mathcal{J}(\boldsymbol{W})\|_\infty \leq \varpi(\kappa) < \infty$，可得：

$$|\widetilde{W}_{ij}^{\mathcal{T}}| \leq \left(\max_{(i,j)}|W_{ij}^{\mathcal{T}-1}| + \frac{\varpi(C_0)}{C}\right).$$

进而得到：

$$\max_{(i,j)}|W_{ij}^{\mathcal{T}-1}| \leq \sqrt{\frac{2}{\alpha_2}\cdot\epsilon_{\mathcal{T}-1}} = C_0,\ \max_{(i,j)}|\widetilde{W}_{ij}^{\mathcal{T}}| \leq \kappa_0.$$

由于$\boldsymbol{W}^{\mathcal{T}} \leftarrow sgn(\widetilde{\boldsymbol{W}}^{\mathcal{T}})\left(\left|\dfrac{\widetilde{\boldsymbol{W}}^{\mathcal{T}}}{1+\frac{\alpha_2}{C}}\right| - \frac{\alpha_1}{C+\alpha_2}\boldsymbol{D}^{\mathcal{T}}\right)_+$，当$(i,j) \in \mathsf{Supp}(\boldsymbol{W}^{\mathcal{T}})$时，有：

$$\begin{aligned}||\boldsymbol{f}_i^{\mathcal{T}} - \boldsymbol{f}_{d+j}^{\mathcal{T}}||^2 = D_{i,j}^{\mathcal{T}} &< \frac{C}{\alpha_1}\cdot\left|\widetilde{\boldsymbol{W}}_{i,j}^{\mathcal{T}}\right| \leq \frac{C}{\alpha_1}\cdot\left(\max_{(i,j)}|W_{ij}^{\mathcal{T}-1}| + \frac{\varpi(C_0)}{C}\right)\\ &\leq \frac{C}{\alpha_1}\cdot\left(C_0 + \frac{\varpi(C_0)}{C}\right) = \delta_1.\end{aligned}$$

(b)的证明：

由(b)，当$\dfrac{C}{\alpha_1}\delta_0 > D_{i,j}^{\mathcal{T}}$时，$\left|\dfrac{\widetilde{\boldsymbol{W}}_{i,j}^{\mathcal{T}}}{1+\frac{\alpha_2}{C}}\right| > \frac{\alpha_1}{C+\alpha_2}D_{i,j}^{\mathcal{T}}$。根据算法6，$|\boldsymbol{W}_{i,j}^{\mathcal{T}}| > 0$成立。 □

引理 C.3. (Yin 等, 2018, 引理1) 令$\boldsymbol{X}$与$\boldsymbol{Y}$为$\mathbb{R}^{n\times n}$的两个正交矩阵。令$\boldsymbol{X} = [\boldsymbol{X}_0, \boldsymbol{X}_1]$且$\boldsymbol{Y} = [\boldsymbol{Y}_0, \boldsymbol{Y}_1]$，其中 $\boldsymbol{X}_0$和$\boldsymbol{Y}_0$分别为$\boldsymbol{X}$和$\boldsymbol{Y}$的第K列，则有：

$$\|\boldsymbol{X}_0\boldsymbol{X}_0^\top - \boldsymbol{Y}_0\boldsymbol{Y}_0^\top\|_F \leq \sqrt{2}\|\boldsymbol{X}_0^\top\boldsymbol{Y}_1\|_F. \tag{C.65}$$

引理 C.4 (sine Θ). (Yu 等, 2014, 定理1) 令$\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}$均为对称方阵，且特征值分别为$\lambda_1 \geq \cdots, \lambda_p$和$\hat{\lambda_1}\cdots\hat{\lambda_p}$。固定$1 \leq K \leq p$，且令 $\boldsymbol{X}_0 = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_K] \in \mathbb{R}^{p\times K}$，$\hat{\boldsymbol{Y}_0} = [\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \cdots, \hat{\boldsymbol{v}}_K]$，$\boldsymbol{X}_1 = [\boldsymbol{v}_{K+1}, \cdots, \boldsymbol{v}_p]$，$\boldsymbol{Y}_1 = [\hat{\boldsymbol{v}}_{K+1}, \cdots, \hat{\boldsymbol{v}}_p]$。对于$1 \leq j \leq p$，有$\boldsymbol{\Sigma}\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j$和$\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{v}}_j = \hat{\lambda}_j\hat{\boldsymbol{v}}_j$。若$\delta = |\hat{\lambda}_{K+1} - \lambda_K| > 0$，则有：

$$\|\boldsymbol{X}_0^\top\boldsymbol{Y}_1\|_F \leq \frac{\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F}{\delta} \tag{C.66}$$

证明. $\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}}$和$\mathcal{L}^{\star}_{\mathcal{G}_{BI}}$分别为二部图$\mathcal{G}^{\mathcal{T}}$和$\mathcal{G}^{\star}$对应的图拉普拉斯矩阵。令 $\boldsymbol{v}^{\mathcal{T}}_1, \cdots, \boldsymbol{v}^{\mathcal{T}}_k$和$\boldsymbol{v}^{\star}_1, \cdots, \boldsymbol{v}^{\star}_k$分别为$\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}}$和$\mathcal{L}^{\star}_{\mathcal{G}_{BI}}$最小的$k$个特征根，且有$\boldsymbol{V}^{\mathcal{T}}_k = [\boldsymbol{v}^{\mathcal{T}}_1 \cdots, \boldsymbol{v}^{\mathcal{T}}_k]$，$\boldsymbol{V}^{\star}_k = [\boldsymbol{v}^{\mathcal{T}}_1 \cdots, \boldsymbol{v}^{\mathcal{T}}_k]$，$\boldsymbol{U}^{\mathcal{T}} = \boldsymbol{V}^{\mathcal{T}}_k \boldsymbol{V}^{\mathcal{T}^{\top}}_k$，$\boldsymbol{U}^{\star} = \boldsymbol{V}^{\star}_k \boldsymbol{V}^{\star^{\top}}_k$。根据引理C.3和引理C.4有:

$$\max_{(i,j)} |\boldsymbol{U}^{\mathcal{T}}_{i,j} - \boldsymbol{U}^{\star}_{i,j}| \ \leq \ ||\boldsymbol{U}^{\mathcal{T}} - \boldsymbol{U}^{\star}||_F \leq \frac{\sqrt{2}||\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}} - \mathcal{L}^{\star}_{\mathcal{G}_{BI}}||_F}{\lambda_{k+1}(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}})} \tag{C.67}$$

因为$\boldsymbol{W}^{\mathcal{T}} \in \mathcal{H}_{C_0}$，且$\boldsymbol{W}^{\star} \in \mathcal{H}_{C_0}$，所以：

$$||\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}} - \mathcal{L}^{\star}_{\mathcal{G}_{BI}}||_F \ \leq \ (\sqrt{d+T} + \sqrt{2}) \cdot ||\boldsymbol{W}^{\mathcal{T}} - \boldsymbol{W}^{\star}||_F \ \leq \ 2 \cdot (\sqrt{d+T} + \sqrt{2}) \cdot C_0 \tag{C.68}$$

根据谱嵌入的定义，有：

$$\max_{(i,j)} |D^{\mathcal{T}}_{i,j} - D^{\star}_{i,j}| \leq 4 \cdot \max_{(i,j)} |\boldsymbol{U}^{\mathcal{T}}_{i,j} - \boldsymbol{U}^{\star}_{i,j}| \ \leq \ 8\sqrt{2} \cdot (\sqrt{d+T} + \sqrt{2}) \cdot \frac{C_0}{\lambda_{k+1}(\mathcal{L}^{\mathcal{T}}_{\mathcal{G}_{BI}})} = 8\sqrt{2}\xi. \tag{C.69}$$

同理可知：

$$D^{\mathcal{T}}_{i,j} \in [D^{\star}_{i,j} - 8\sqrt{2}\xi, D^{\star}_{i,j} + 8\sqrt{2}\xi] \tag{C.70}$$

由于$\mathcal{G}^{\star}_{BI}$包含k个连通分量，故有

$$D^{\star}_{i,j} = \begin{cases} \dfrac{1}{n_{\mathcal{G}(i)}} + \dfrac{1}{n_{\mathcal{G}(j)}} \ \geq \ \dfrac{1}{n^{\uparrow}_1} + \dfrac{1}{n^{\uparrow}_2} \ = \ \beta, & \mathcal{G}^{\star}(i) \neq \mathcal{G}^{\star}(j) \\ 0, & \text{otherwise.} \end{cases}$$

根据假设有：

$$\min_{(i,j):\mathcal{G}^{\star}(i) \neq \mathcal{G}^{\star}(j)} D^{\mathcal{T}}_{i,j} \geq \beta - 8\sqrt{2}\xi > 8\sqrt{2}\xi \geq \max_{(i,j):\mathcal{G}^{\star}(i) = \mathcal{G}^{\star}(j)} D^{\mathcal{T}}_{i,j}. \tag{C.71}$$

针对(a)，由于$8\sqrt{2}\xi < \delta_1 < \beta - 8\sqrt{2}\xi$，故：

$$\left\{(i,j) : ||\boldsymbol{f}^{\mathcal{T}}_i - \boldsymbol{f}^{\mathcal{T}}_{d+j}||^2_2 < \delta_1\right\} = \left\{(i,j) : \mathcal{G}(i) = \mathcal{G}(j)\right\}$$

针对(b)，由于$8\sqrt{2}\xi < \min\{\delta_1, \delta_2\} \leq \max\{\delta_1, \delta_2\} < \beta - 8\sqrt{2}\xi$，故有：

$$\left\{(i,j) : ||\boldsymbol{f}^{\mathcal{T}}_i - \boldsymbol{f}^{\mathcal{T}}_{d+j}||^2_2 < \delta_1\right\} = \left\{(i,j) : ||\boldsymbol{f}^{\mathcal{T}}_i - \boldsymbol{f}^{\mathcal{T}}_{d+j}||^2_2 < \delta_2\right\} = \left\{(i,j) : \mathcal{G}(i) = \mathcal{G}(j)\right\}.$$

由引理C.2，证毕。 □ □

C.5 个性化属性预测模型的优化方法

C.5.1 收敛性分析

下面证明$\mathcal{J}=\sum_i \ell_i$的梯度是Lipschitz连续的，且相对于$\boldsymbol{\Theta_g}$的偏微分有界。注意下文中交替使用了三种等价的经验损失表达形式：$\mathcal{J}$，$\mathcal{J}(\boldsymbol{\Theta})$和$\mathcal{J}(\boldsymbol{\Theta}_c,\boldsymbol{\Theta}_g,\boldsymbol{\Theta}_p)$，其中$\boldsymbol{\Theta}=[\boldsymbol{\Theta}_c;vec(\boldsymbol{\Theta}_g);vec(\boldsymbol{\Theta}_p)]$。

引理 C.5. 若数据满足如下有界条件：

$$\forall i,\ \|\boldsymbol{X}^{(i)}\|_2=\vartheta_{X_i}<\infty,\ n_{+,i}\geq 1,\ n_{-,i}\geq 1.$$

(1) 给定任意两个不同的参数$\boldsymbol{W},\boldsymbol{W}'$，有：

$$\|\nabla\mathcal{J}(\boldsymbol{\Theta})-\nabla\mathcal{J}(\boldsymbol{\Theta}')\|_F\leq\varrho_{\boldsymbol{\Theta}}\Delta\boldsymbol{\Theta}$$

(2) 对于任意$\infty>\xi_c>0,\infty>\xi_g>0$，$\infty>\xi_p>0$，有：

$$\sup_{\|\boldsymbol{\Theta}_c\|_2\leq\xi_c,\|\boldsymbol{\Theta}_g\|_F\leq\xi_g,\|\boldsymbol{\Theta}_p\|_F\leq\xi_p}\|\nabla_{\boldsymbol{\Theta_g}}\mathcal{J}\|_F\leq\varkappa(\xi_c,\xi_g,\xi_p)\tag{C.72}$$

其中$\boldsymbol{\Theta}=[\boldsymbol{\Theta}_c;vec(\boldsymbol{\Theta}_g);vec(\boldsymbol{\Theta}_p)]$，$\boldsymbol{\Theta}'=[\boldsymbol{\Theta}';vec(\boldsymbol{\Theta}'_g);vec(\boldsymbol{\Theta}'_p)]$，$\Delta\boldsymbol{\Theta}=\|\boldsymbol{\Theta}-\boldsymbol{\Theta}'\|$，$\varrho_{\boldsymbol{\Theta}}=3T\sqrt{(2T+1)}\max_i\left\{\frac{n_i\vartheta_{X_i}^2}{n_{+,i}n_{-,i}}\right\}$，
$\varkappa(\xi_c,\xi_g,\xi_p)=\frac{n_i\vartheta_{X_i}}{\sqrt{n_{+,i}}n_{-.i}}\sum_{i=1}^{T}\left((\xi_c+\xi_g+\xi_p)\frac{\vartheta_{X_i}}{\sqrt{n_{+,i}}}+1\right)$。

证明. **(1)的证明**

记

$$\begin{aligned}
dL_i&=\boldsymbol{X}^{(i)\top}\mathcal{L}^{(i)}_{AUC}\boldsymbol{X}^{(i)}(\boldsymbol{W}^{(i)}-\boldsymbol{W}'^{(i)}),\\
d_c&=\nabla_{\boldsymbol{\Theta}_c}\mathcal{J}(\boldsymbol{\Theta})-\nabla_{\boldsymbol{\Theta}'_c}\mathcal{J}(\boldsymbol{\Theta}'),\\
d_g^{(i)}&=\nabla_{\boldsymbol{\Theta}_g^{(i)}}\mathcal{J}(\boldsymbol{\Theta})-\nabla_{\boldsymbol{\Theta}'^{(i)}_g}\mathcal{J}(\boldsymbol{\Theta}'),\\
d_p^{(i)}&=\nabla_{\boldsymbol{\Theta}_p^{(i)}}\mathcal{J}(\boldsymbol{\Theta})-\nabla_{\boldsymbol{\Theta}'^{(i)}_p}\mathcal{J}(\boldsymbol{\Theta}').
\end{aligned}$$

注意到：

$$\boldsymbol{W}^{(i)}=\boldsymbol{\Theta}_c+\boldsymbol{\Theta}_g^{(i)}+\boldsymbol{\Theta}_p^{(i)},\ \boldsymbol{W}'^{(i)}=\boldsymbol{\Theta}'_c+\boldsymbol{\Theta}'^{(i)}_g+\boldsymbol{\Theta}'^{(i)}_p$$

因此有：

$$
\begin{aligned}
&\|\nabla\mathcal{J}(\boldsymbol{\Theta}) - \nabla\mathcal{J}(\boldsymbol{\Theta}')\| \\
=\ &\left(\|d_c\|^2 + \sum_i \|d_g^{(i)}\|^2 + \sum_i \|d_p^{(i)}\|^2\right)^{1/2} \\
\leq\ &\|d_c\| + \sum_i \|d_g^{(i)}\| + \sum_i \|d_p^{(i)}\| \\
=\ &\|\sum_{i=1}^{T} dL_i\| + 2\sum_{i=1}^{T}\|dL_i\| \\
\leq\ &3C_{max}\sum_{i=1}^{T}\|\boldsymbol{W}^{(i)} - \boldsymbol{W}'^{(i)}\| \\
\leq\ &3C_{max}T\left(\|\boldsymbol{\Theta}_c - \boldsymbol{\Theta}_c'\| + \sum_{i=1}^{T}\|\boldsymbol{\Theta}_g^{(i)} - \boldsymbol{\Theta}_g^{(i)\prime}\| + \sum_{i=1}^{T}\|\boldsymbol{\Theta}_p{}^{(i)} - \boldsymbol{\Theta}_p{}^{(i)\prime}\|\right) \\
\leq\ &3C_{max}T\sqrt{2T+1}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|
\end{aligned}
\tag{C.73}
$$

其中，

$$
C_{max} = \max_i \left(\|\boldsymbol{X}^{(i)\top}\mathcal{L}_{AUC}^{(i)}\boldsymbol{X}^{(i)}\|_2\right).
$$

为证明上述不等式，有：

$$
\forall i, \|\boldsymbol{X}^{(i)\top}\mathcal{L}_{AUC}^{(i)}\boldsymbol{X}^{(i)}\|_2 \leq \|\boldsymbol{X}^{(i)}\|_2^2\|\mathcal{L}_{AUC}^{(i)}\|_2. \tag{C.74}
$$

且根据(Zhang, 2011)中定理3.3可得：

$$
\|\mathcal{L}_{AUC}^{(i)}\|_2 = \frac{n_i}{n_{+,i}n_{-,i}}. \tag{C.75}
$$

综合C.74和C.75上述不等式成立，(1)证毕。

(2)的证明

省略上标，有：

$$
\begin{aligned}
\sup\|\nabla_{\boldsymbol{\Theta}_g}\mathcal{J}\|_F &\leq \sum_{i=1}^{T}\sup\|\nabla_{\boldsymbol{\Theta}_g{}^{(i)}}\mathcal{J}\|_2 \\
&\leq \sum_{i=1}^{T}\sup\|\boldsymbol{X}^{(i)\top}\mathcal{L}_{AUC}^{(i)}\boldsymbol{X}^{(i)}\|_2\cdot\|\boldsymbol{W}^{(i)}\|_2 + \|\boldsymbol{X}^{(i)\top}\mathcal{L}_{AUC}^{(i)}\|_2\cdot\|\widetilde{\boldsymbol{y}}^{(i)}\|_2 \\
&\leq \frac{n_i\vartheta_{X_i}}{\sqrt{n_{+,i}}n_{-.i}}\sum_{i=1}^{T}\left((\xi_c+\xi_g+\xi_p)\frac{\vartheta_{X_i}}{\sqrt{n_{+,i}}} + 1\right) = \varkappa(\xi_c,\xi_g,\xi_p).
\end{aligned}
\tag{C.76}
$$

□

同问题$(\boldsymbol{P})$相似，此处定义$(\boldsymbol{Q})$的替代问题，且命名为$(\boldsymbol{Q}^{\star})$：

$$(\boldsymbol{Q}^{\star}) \min_{\boldsymbol{\Theta},\boldsymbol{U}\in\Gamma} \mathcal{J}(\boldsymbol{\Theta}_c,\boldsymbol{\Theta}_g,\boldsymbol{\Theta}_p)+\frac{\alpha_1}{2}\|\boldsymbol{\Theta}_c\|_2^2+\alpha_2\langle\boldsymbol{\Theta}_{\mathcal{G}},\boldsymbol{U}\rangle+\frac{\alpha_3}{2}\|\boldsymbol{\Theta}_g\|_F^2+\alpha_4\|\boldsymbol{\Theta}_p\|_{1,2}+\frac{\alpha_5}{2}\|\boldsymbol{U}\|_F^2. \tag{C.77}$$

此外，记替代目标函数为：

$$\mathcal{F}(\boldsymbol{\Theta}_c,\boldsymbol{\Theta}_g,\boldsymbol{\Theta}_p,\boldsymbol{U}) = \widetilde{\mathcal{F}}(\boldsymbol{\Theta}_c,\boldsymbol{\Theta}_g,\boldsymbol{\Theta}_p,\boldsymbol{U}) + \frac{\alpha_5}{2}\|\boldsymbol{U}\|_F^2.$$

针对该问题，显然存在与算法7相似的解法，且具有以下性质：

引理C.6. 记$\mathcal{F}_t = \mathcal{F}(\boldsymbol{\Theta}_c^t,\boldsymbol{\Theta}_g^t,\boldsymbol{\Theta}_p^t,\boldsymbol{U}^t)$，$(\boldsymbol{\Theta}_c^t,\boldsymbol{\Theta}_g^t,\boldsymbol{\Theta}_p^t,\boldsymbol{U}^t)$为第$t$轮迭代的参数，取$C > \varrho_{\Theta}$，$0 < \alpha_5 < 2C\min_t \check{\delta}(\mathcal{L}^t_{\mathcal{G}_{BI}}) < +\infty$，则以下性质成立：

(1) 当以下条件满足时，序列$\{\mathcal{F}_t\}$非递增：

$$\mathcal{F}_{t+1} \le \mathcal{F}_t - \min\left\{\frac{C-\varrho_{\Theta}}{2},\frac{\alpha_5}{2}\right\}\cdot\left(\|\Delta(\boldsymbol{\Theta}_c^t)\|_F^2+\|\Delta(\boldsymbol{\Theta}_g^t)\|_F^2+\|\Delta(\boldsymbol{\Theta}_p^t)\|_F^2+\|\Delta(\boldsymbol{U}^t)\|_F^2\right),$$

其中 $\Delta(\boldsymbol{\Theta}_c^t) = \boldsymbol{\Theta}_c^{t+1} - \boldsymbol{\Theta}_c^t$，$\Delta(\boldsymbol{\Theta}_g^t) = \boldsymbol{\Theta}_g^{t+1} - \boldsymbol{\Theta}_g^t$，$\Delta(\boldsymbol{\Theta}_p^t) = \boldsymbol{\Theta}_p^{t+1} - \boldsymbol{\Theta}_p^t$，$\Delta(\boldsymbol{U}^t) = \boldsymbol{U}^{t+1} - \boldsymbol{U}^t$。

(2) $\sum_{i=1}^{\infty}\|\Delta(\boldsymbol{\Theta}_c^t)\|_F^2 + \|\Delta(\boldsymbol{\Theta}_g^t)\|_F^2 + \|\Delta(\boldsymbol{\Theta}_p^t)\|_F^2 + \|\Delta(\boldsymbol{U}^t)\|_F^2 < \infty$。进一步，有：$\lim_{t\to\infty}||\Delta(\boldsymbol{\Theta}_c^t)|| = 0$ $\lim_{t\to\infty}||\Delta(\boldsymbol{\Theta}_g^t)|| = 0$ $\lim_{t\to\infty}||\Delta(\boldsymbol{\Theta}_p^t)|| = 0$ $\lim_{t\to\infty}||\Delta(\boldsymbol{U}^t)|| = 0$。

(3) 存在$\{\boldsymbol{\Theta}_c{}^{k_j},\boldsymbol{\Theta}_g{}^{k_j},\boldsymbol{\Theta}_p{}^{k_j},\boldsymbol{U}^{k_j}\}$的子序列，存在聚点$\{\boldsymbol{\Theta}_c{}^{\star},\boldsymbol{\Theta}_g{}^{\star},\boldsymbol{\Theta}_p{}^{\star},\boldsymbol{U}^{\star}\}$使得：

$$\begin{aligned}\{\boldsymbol{\Theta}_c{}^{k_j},\ \boldsymbol{\Theta}_g{}^{k_j},\ \boldsymbol{\Theta}_p{}^{k_j},\ \boldsymbol{U}^{k_j}\} &\to \{\boldsymbol{\Theta}_c{}^{\star},\ \boldsymbol{\Theta}_g{}^{\star},\ \boldsymbol{\Theta}_p{}^{\star},\ \boldsymbol{U}^{\star}\},\\ \mathcal{F}(\boldsymbol{\Theta}_c{}^{k_j},\ \boldsymbol{\Theta}_g{}^{k_j},\ \boldsymbol{\Theta}_p{}^{k_j},\ \boldsymbol{U}^{k_j}) &\to \mathcal{F}(\boldsymbol{\Theta}_c{}^{\star},\ \boldsymbol{\Theta}_g{}^{\star},\ \boldsymbol{\Theta}_p{}^{\star},\ \boldsymbol{U}^{\star}).\end{aligned} \tag{C.78}$$

(4) $\mathcal{F}(\cdot,\cdot,\cdot,\cdot)$为$KL$函数。

(5) 次梯度满足：

$$dist(\boldsymbol{0},\partial_{\Theta}\mathcal{F}(\boldsymbol{\Theta}_c^t,\boldsymbol{\Theta}_g^t,\boldsymbol{\Theta}_p^t,\boldsymbol{U}^t)) \le \left[3\cdot(C+\varrho_{\Theta})+\alpha_1(\sqrt{d}+\sqrt{T}+2)\right]\cdot\|\boldsymbol{\Theta}^t-\boldsymbol{\Theta}^{t-1}\|. \tag{C.79}$$

证明. 由于$\nabla\mathcal{J}$为ϱ_{Θ}-Lipschitz连续，以下结论成立：

$$\begin{aligned}\mathcal{J}_{t+1} \le& \mathcal{J}_t + \left\langle\nabla_{\boldsymbol{\Theta}_c}\mathcal{J}_t,\Delta\left(\boldsymbol{\Theta}_c^t\right)\right\rangle + \left\langle\nabla_{\boldsymbol{\Theta}_g}\mathcal{J}_t,\Delta\left(\boldsymbol{\Theta}_g^t\right)\right\rangle + \left\langle\nabla_{\boldsymbol{\Theta}_p}\mathcal{J}_t,\Delta\left(\boldsymbol{\Theta}_p^t\right)\right\rangle\\ &+\frac{\varrho_{\Theta}}{2}\cdot\left[\|\Delta(\boldsymbol{\Theta}_c^t)\|_F^2+\|\Delta(\boldsymbol{\Theta}_g^t)\|_F^2+\|\Delta(\boldsymbol{\Theta}_p^t)\|_F^2\right].\end{aligned} \tag{C.80}$$

由式 (C.80)和子问题的强凸性，(1)的证明与引理C.1相似。由每个子问题的最优条件，(2)的证明与引理C.1中(2)的证明相似。根据波尔查诺-魏尔斯特拉斯定理、每个子问题的最优条件，以及损失函数连续且下界远大于0可知，(3)的证明与引理C.1中(3)的证明相似。由于损失函数$\mathcal{J}$为$\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_g, \boldsymbol{\Theta}_p$的多项式函数，故$\mathcal{J}$正定。类似地，$\|\boldsymbol{\Theta}_c\|_2^2$是可定义的。由(Lau 等, 2018)可知$||\boldsymbol{\Theta}_g||_{1,2}$同样是可定义的。综合引理C.1中证明，可知$\mathcal{F}(\cdot,\cdot,\cdot,\cdot)$是可定义的且为KL函数，由此(4)证毕。 □

以下给出定理5.7的证明：

证明. 证明过程同5.4，可由引理C.6、及定理5.5得出。 □

最后，可将定理5.6的证明扩展至定理5.8：

证明. 证明过程同引理C.5-(2)、定理5.6相似。唯一差别为由于$\widetilde{F}_t$非增，从而：

$$\|\boldsymbol{\Theta}_c^t\|_F \leq \sqrt{\frac{\widetilde{F}_0}{\alpha_3}},\ \|\boldsymbol{\Theta}_g^t\|_F \leq \sqrt{\frac{\widetilde{F}_0}{\alpha_2}},\ \|\boldsymbol{\Theta}_p^t\|_F \leq \|\boldsymbol{\Theta}_p^t\|_{1,2} \leq 2 \cdot \frac{\widetilde{F}_0}{\alpha_4}. \tag{C.81}$$

□

参考文献

Agarwal S. Surrogate regret bounds for bipartite ranking via strongly proper losses [J]. Journal of Machine Learning Research, 2014, 15(1): 1653-1674.

Agarwal S, Graepel T, Herbrich R, et al. Generalization bounds for the area under the roc curve [J]. Journal of Machine Learning Research, 2005, 6(Apr): 393-425.

Alan A H, Raskutti B. Optimising area under the roc curve using gradient descent [J]. International Conference on Machine Learning, 2004: 49-56.

Andreani R, Haeser G, Viana D S. Optimality conditions and global convergence for nonlinear semidefinite programming [J]. Mathematical Programming, 2020, 180(1): 203-235.

Argyriou A, Evgeniou T, Pontil M. Convex multi-task feature learning [J]. Machine Learning, 2008, 73(3): 243-272.

Argyriou A, Pontil M, Ying Y, et al. A spectral regularization framework for multi-task structure learning [C]//Neurips. 2008b: 25-32.

Attouch H, Bolte J, Redont P, et al. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality [J]. Mathematics of Operations Research, 2010, 35(2): 438-457.

Attouch H, Bolte J, Svaiter B F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods [J]. Mathematical Programming, 2013, 137(1-2): 91-129.

Bach F R, Jordan M I. Learning spectral clustering [C]//NeurIPS. 2004: 305-312.

Bakker B, Heskes T. Task clustering and gating for bayesian multitask learning [J]. Journal of Machine Learning Research, 2003, 4(May): 83-99.

Bartlett P L, Mendelson S. Rademacher and gaussian complexities: Risk bounds and structural results [J]. Journal of Machine Learning Research, 2002, 3(Nov): 463-482.

Barua S, Islam M M, Yao X, et al. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 26(2): 405-425.

Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems [J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202.

Bochnak J, Coste M, Roy M F. Real algebraic geometry: volume 36 [M]. Springer Science & Business Media, 2013.

Bolte J, Daniilidis A, Lewis A, et al. Clarke subgradients of stratifiable functions [J]. SIAM Journal on Optimization, 2007, 18(2): 556-572.

Boucheron S, Lugosi G, Massart P. Concentration inequalities: A nonasymptotic theory of independence [M]. Oxford university press, 2013.

Bowyer K W, Kranenburg C, Dougherty S. Edge detector evaluation using empirical ROC curves [C]//1999 Conference on Computer Vision and Pattern Recognition (CVPR '99), 23-25 June 1999, Ft. Collins, CO, USA. IEEE Computer Society, 1999: 1354-1359.

Boyd S, Vandenberghe L. Convex optimization [M]. Cambridge University Press, 2004.

Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms [J]. Pattern Recognit., 1997, 30(7): 1145-1159.

Calders T, Jaroszewicz S. Efficient auc optimization for classification [J]. European Conference on Principles of Data Mining and Knowledge Discovery, 2007: 42-53.

Cao J, Li Y, Zhang Z. Partially shared multi-task convolutional neural network with local constraint for face attribute learning [C]//CVPR. 2018: 4290-4299.

Cao K, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss [C]//Advances in Neural Information Processing Systems. 2019.

Chen J, Zhou J, Ye J. Integrating low-rank and group-sparse structures for robust multi-task learning [C]//KDD. 2011: 42-50.

Clémençon S, Achab M. Ranking data with continuous labels through oriented recursive partitions [C]//Advances in Neural Information Processing Systems. 2017: 4600-4608.

Clémençon S, Lugosi G, Vayatis N, et al. Ranking and empirical minimization of u-statistics [J]. The Annals of Statistics, 2008, 36(2): 844-874.

Clémençon S, Robbiano S, Vayatis N. Ranking data with ordinal labels: optimality and pairwise aggregation [J]. Machine Learning, 2013, 91(1): 67-104.

Cortes C, Mohri M. Auc optimization vs. error rate minimization [J]. Advances in Neural Information Processing Systems, 2003: 313-320.

Cortes C, Kuznetsov V, Mohri M, et al. Structured prediction theory based on factor graph complexity [J]. Advances in Neural Information Processing Systems, 2016: 2514-2522.

Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9268-9277.

Dodd L E, Pepe M S. Partial auc estimation and regression [J]. Biometrics, 2003, 59(3): 614-623.

Egan J. Signal detection theory and roc analysis. series in cognitition and perception [M]. Academic Press, New York, 1975.

Elhamifar E, Vidal R. Sparse subspace clustering [C]//CVPR. 2009: 2790-2797.

Fan K. On a theorem of Weyl concerning eigenvalues of linear transformations I [J]. PNAS, 1949, 35(11): 652-655.

Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes [C]//CVPR. 2009: 1778-1785.

Favaro P, Vidal R, Ravichandran A. A closed form solution to robust subspace estimation and clustering [C]//CVPR. 2011: 1801-1807.

Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

Ferri C, Hernández-Orallo J, Salido M A. Volume under the ROC surface for multi-class problems [J]. European Conference on Machine Learning, 2003: 108-120.

Freund Y, Schapire R E. Experiments with a new boosting algorithm [C]//Saitta L. International Conference on Machine Learning. Morgan Kaufmann, 1996: 148-156.

Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences [J]. Journal of machine learning research, 2003, 4(Nov): 933-969.

Freund Y, Iyer R D, Schapire R E, et al. An efficient boosting algorithm for combining preferences [J]. J. Mach. Learn. Res., 2003, 4: 933-969.

Fu Y, Liu C, Li D, et al. Parsimonious deep learning: A differential inclusion approach with global convergence [J]. arXiv preprint arXiv:1905.09449, 2019.

Fujino A, Ueda N. A semi-supervised AUC optimization method with generative models [C]//IEEE International Conference on Data Mining. 2016: 883-888.

Gao W, Wang W. Analysis of k-partite ranking algorithm in area under the receiver operating characteristic curve criterion [J]. International Journal of Computer Mathematics, 2018, 95(8): 1527-1547.

Gao W, Zhou Z. On the consistency of AUC pairwise optimization [J]. International Joint Conference on Artificial Intelligence, 2015: 939-945.

Gao W, Wang L, Jin R, et al. One-pass auc optimization [J]. International Conference on Machine Learning, 2013: 906-914.

Gao W, Wang L, Jin R, et al. One-pass AUC optimization [J]. Artif. Intell., 2016, 236: 1-29.

Golowich N, Rakhlin A, Shamir O. Size-independent sample complexity of neural networks [C]// 2018: 297-299.

Grave E, Obozinski G, Bach F R. Trace lasso: a trace norm regularization for correlated designs [C]//NeurIPS. 2011: 2187-2195.

Han H, Wang W Y, Mao B H. Borderline-smote: a new over-sampling method in imbalanced data sets learning [J]. International Conference on Intelligent Computing, 2005: 878-887.

Han L, Zhang Y. Learning multi-level task groups in multi-task learning [C]//AAAI. 2015: 2638-2644.

Han L, Zhang Y. Multi-stage multi-task learning with reduced rank [C]//AAAI. 2016: 1638-1644.

Hand D J, Till R J. A simple generalisation of the area under the roc curve for multiple class classification problems [J]. Machine Learning, 2001, 45(2): 171-186.

Hanley J A, McNeil B J. The meaning and use of the area under a receiver operating characteristic (roc) curve [J]. Radiology, 1982, 143(1): 29-36.

Heskes T. Solving a huge number of similar tasks: A combination of multi-task learning and a hierarchical bayesian approach [C]//ICML. 1998: 233-241.

Honzik P, Kucera P, Hyncica O, et al. Novel method for evaluation of multi-class area under receiver operating characteristic [J]. International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009: 1-4.

Jaskowiak P A, Costa I G, Campello R J G B. The area under the ROC curve as a measure of clustering quality [J]. CoRR, 2020, abs/2009.02400.

Jeong J Y, Jun C H. Variable selection and task grouping for multi-task learning [C]//KDD. 2018: 1589-1598.

Joachims T. A support vector method for multivariate performance measures [J]. International Conference on Machine Learning, 2005: 377-384.

Joachims T. Training linear svms in linear time [J]. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006: 217-226.

Kang Z, Grauman K, Sha F. Learning with whom to share in multi-task feature learning. [C]//ICML. 2011: 521-528.

Kingma D P, Ba J. Adam: A method for stochastic optimization [C]//3rd International Conference on Learning Representations. 2015.

Korolyuk V S, Borovskich Y V. Theory of u-statistics: volume 273 [M]. Springer Science & Business Media, 2013.

Kovashka A, Grauman K. Discovering attribute shades of meaning with the crowd [J]. International Journal of Computer Vision, 2015, 114(1): 56-73.

Kovashka A, Parikh D, Grauman K. Whittlesearch: Image search with relative attribute feedback [C]//CVPR. 2012: 2973-2980.

Kumar A, III H D. Learning task grouping and overlap in multi-task learning [C]//ICML. 2012: 1723-1730.

Lau T T K, Zeng J, Wu B, et al. A proximal block coordinate descent algorithm for deep neural network training [C]//ICLR Workshop. 2018.

Ledoux M, Talagrand M. Probability in banach spaces: isoperimetry and processes [M]. Springer Science & Business Media, 2013.

Lee G, Yang E, Hwang S. Asymmetric multi-task learning based on task relatedness and loss [C]// ICML. 2016: 230-238.

Lemaître G, Nogueira F, Aridas C K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning [J]. Journal of Machine Learning Research, 2017, 18 (1): 559-563.

Li C G, Vidal R. Structured sparse subspace clustering: A unified optimization framework [C]// CVPR. 2015a: 277-286.

Li H, Lin Z. Accelerated proximal gradient methods for nonconvex programming [J]. Advances in Neural Information Processing Systems, 2015: 379-387.

Li Y, Fu K, Wang Z, et al. Multi-task representation learning for travel time estimation [C]//KDD. 2018: 1695-1704.

Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]//IEEE international conference on computer vision. 2017a: 2980-2988.

Lin Y, Yang L, Lin Z, et al. Factorization for projective and metric reconstruction via truncated nuclear norm [C]//IJCNN. 2017b: 470-477.

Lin Y, Yang S, Stoyanov V, et al. A multi-lingual multi-task architecture for low-resource sequence labeling [C]//ACL. 2018: 799-809.

Liu G, Lin Z, Yan S, et al. Robust recovery of subspace structures by low-rank representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 171.

Liu P, Qiu X, Huang X. Adversarial multi-task learning for text classification [C]//ACL. 2017a: 1-10.

Liu R, Cheng S, Liu X, et al. A bridging framework for model optimization and deep propagation [C]//NIPS. 2018.

Liu R, Cheng S, Ma L, et al. Deep proximal unrolling: Algorithmic framework, convergence analysis and applications [J]. IEEE Transactions on Image Processing, 2019, 28(10): 5013-5026.

Liu R, Zhang Y, Cheng S, et al. A theoretically guaranteed deep optimization framework for robust compressive sensing mri [J]. AAAI Conference on Artificial Intelligence, 2019: 4368-4375.

Liu S, Pan S J. Adaptive group sparse multi-task learning via trace lasso. [C]//IJCAI. 2017b: 2358-2364.

Long P M, Sedghi H. Generalization bounds for deep convolutional neural networks [C]// International Conference on Learning Representations. 2020.

Lu C, Zhu C, Xu C, et al. Generalized singular value thresholding [C]//AAAI. 2015: 1805-1811.

Lu C, Feng J, Lin Z, et al. Nonconvex sparse spectral clustering by alternating direction method of multipliers and its convergence analysis [C]//AAAI. 2018: 3714-3721.

Lu C, Feng J, Lin Z, et al. Subspace clustering by block diagonal representation [J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(2): 487-501.

Luo C, Li Z, Huang K, et al. Zero-shot learning via attribute regression and class prototype rectification [J]. IEEE Trans. Image Processing, 2018, 27(2): 637-648.

Lyu S, Ying Y. A univariate bound of area under ROC [J]. Conference on Uncertainty in Artificial Intelligence, 2018: 43-52.

Mani I, Zhang I. knn approach to unbalanced data distributions: a case study involving information extraction [J]. ICML Workshop on Learning from Imbalanced Datasets, 2003, 126.

Maurer A. A vector-contraction inequality for rademacher complexities [J]. International Conference on Algorithmic Learning Theory, 2016: 3-17.

Maurer A, Pontil M. Uniform concentration and symmetrization for weak interactions [C]// Conference on Learning Theory. 2019: 2372-2387.

Maurer A, Pontil M. Estimating weighted areas under the ROC curve [C]//Advances in Neural Information Processing Systems. 2020.

Maurer A, Pontil M, Romera-Paredes B. Sparse coding for multitask and transfer learning [C]// ICML. 2013: 343-351.

McDiarmid C. Concentration [M]//Springer, 1998: 195-248.

McDonald A M, Pontil M, Stamos D. New perspectives on k-support and cluster norms [J]. Journal of Machine Learning Research, 2016, 17(155): 1-38.

Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning [J]. 2018.

Mossman D. Three-way rocs [J]. Medical Decision Making, 1999, 19(1): 78-89.

Mozer M C, Dodier R H, Colagrosso M D, et al. Prodding the ROC curve: Constrained optimization of classifier performance [C]//Advances in Neural Information Processing Systems. 2001: 1409-1415.

Narasimhan H, Agarwal S. A structural SVM based approach for optimizing partial AUC [J]. International Conference on Machine Learning, 2013: 516-524.

Narasimhan H, Agarwal S. $\text{Svm}_{\text{pauc}}^{\text{tight}}$: a new support vector method for optimizing partial AUC based on a tight convex upper bound [C]//International Conference on Knowledge Discovery and Data Mining. 2013b: 167-175.

Narasimhan H, Agarwal S. A structural SVM based approach for optimizing partial AUC [C]// International Conference on Machine Learning. 2013c: 516-524.

Narasimhan H, Agarwal S. Support vector algorithms for optimizing the partial area under the ROC curve [J]. Neural Computation, 2017, 29(7): 1919-1963.

Natole M, Ying Y, Lyu S. Stochastic proximal algorithms for auc maximization [J]. International Conference on Machine Learning, 2018: 3707-3716.

Natole M A, Ying Y, Lyu S. Stochastic auc optimization algorithms with linear convergence [J]. Frontiers in Applied Mathematics and Statistics, 2019, 5: 30.

Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm [C]//NeurIPS. 2002: 849-856.

Ni K, Carin L, Dunson D B. Multi-task learning for sequential data via ihmms and the nested dirichlet process [C]//ICML. 2007: 689-696.

Nie F, Huang H, Cai X, et al. Efficient and robust feature selection via joint l2,1-norms minimization [C]//In NIPS. 2010: 1813-1821.

Nie F, Wang X, Huang H. Clustering and projected clustering with adaptive neighbors [C]//KDD. 2014: 977-986.

Nie F, Wang X, Jordan M I, et al. The constrained laplacian rank algorithm for graph-based clustering [C]//AAAI. 2016.

Nie F, Hu Z, Li X. Calibrated multi-task learning [C]//KDD. 2018: 2012-2021.

Oliveira S H G, Goncalves A R, Von Zuben F J. Group lasso with asymmetric structure estimation for multi-task learning [C]//IJCAI. 2019: 3202-3208.

Patterson G, Hays J. Sun attribute database: Discovering, annotating, and recognizing scene attributes [C]//CVPR. IEEE, 2012: 2751-2758.

Popescu P G, Dragomir S S, Sluşanschi E I, et al. Bounds for kullback-leibler divergence [J]. Electronic Journal of Differential Equations, 2016, 2016.

Provost F J, Domingos P M. Tree induction for probability-based ranking [J]. Machine Learning, 2003, 52(3): 199-215.

Qi Y, Liu D, Dunson D B, et al. Multi-task compressive sensing with dirichlet process priors [C]// ICML. 2008: 768-775.

Rajaram S, Agarwal S. Generalization bounds for k-partite ranking [C]//NIPS Workshop on Learning to Rank. 2005: 18-23.

Ralaivola L, Szafranski M, Stempfel G. Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes [J]. Journal of Machine Learning Research, 2010, 11(Jul): 1927-1956.

Reeve H W, Kaban A. Optimistic bounds for multi-output prediction [J]. arXiv preprint arXiv:2002.09769, 2020.

Rockafellar R T, Wets R J B. Variational analysis: volume 317 [M]. Springer Science & Business Media, 2009.

Sadovnik A, Gallagher A C, Parikh D, et al. Spoken attributes: Mixing binary and relative attributes to say the right thing [C]//ICCV. 2013: 2160-2167.

Sakai T, Niu G, Sugiyama M. Semi-supervised AUC optimization based on positive-unlabeled learning [J]. Machine Learning, 2018, 107(4): 767-794.

Sammut C, Webb G I. Encyclopedia of machine learning [M]. Springer Science & Business Media, 2011.

Schapire R E, Freund Y, Bartlett P, et al. Boosting the margin: A new explanation for the effectiveness of voting methods [J]. Annals of statistics, 1998, 26(5): 1651-1686.

Shen S Q, Yang B B, Gao W. Auc optimization with a reject option [C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. 2020: 5684-5691.

Smith M R, Martinez T, Giraud-Carrier C. An instance level analysis of data complexity [J]. Machine Learning, 2014, 95(2): 225-256.

Song F, Tan X, Chen S. Exploiting relationship between attributes for improved face verification [J]. Computer Vision and Image Understanding, 2014, 122: 143-154.

Sra S, Nowozin S, Wright S J. Optimization for machine learning [M]. Mit Press, 2012.

Su C, Zhang S, Yang F, et al. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping [J]. Pattern Recognition, 2017, 66: 4-15.

Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [C]//CVPR. 2016: 2818-2826.

Thrun S, O'Sullivan J. Discovering structure in multiple learning tasks: The TC algorithm [C]// ICML. 1996: 489-497.

Tibshirani R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

Tomek I. Two modifications of cnn [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1976, SMC-6(11): 769-772.

Uematsu K, Lee Y. Statistical optimality in multipartite ranking and ordinal regression [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(5): 1080-1094.

Usunier N, Amini M R, Gallinari P. A data-dependent generalisation error bound for the auc [J]. ICML Workshop on ROC Analysis in Machine Learning, 2005.

Usunier N, Amini M R, Gallinari P. Generalization error bounds for classifiers trained with interdependent data [J]. Advances in Neural Information Processing Systems, 2006: 1369-1376.

Van den Dries L, Miller C, et al. Geometric categories and o-minimal structures [J]. Duke Math. J, 1996, 84(2): 497-540.

Walter S D. The partial area under the summary roc curve [J]. Statistics in medicine, 2005, 24(13): 2025-2040.

Wang S, Minku L L. Auc estimation and concept drift detection for imbalanced data streams with multiple classes [C]//2020 International Joint Conference on Neural Networks. 2020: 1-8.

Wang S, Yuan X, Yao T, et al. Efficient subspace segmentation via quadratic programming. [C]// AAAI. 2011: 519-524.

Wang Y, Kwok J T, Yao Q, et al. Zero-shot learning with a partial set of observed attributes [C]// IJCNN. 2017: 3777-3784.

Wang Y, Wipf D P, Ling Q, et al. Multi-task learning for subspace segmentation [C]//ICML. 2015: 1209-1217.

Wipf D P, Dong Y, Xin B. Subspace clustering with a twist. [C]//UAI. 2016.

Woods K S, Bowyer K W. Generating ROC curves for artificial neural networks [J]. IEEE Transactions Medical Imaging, 1997, 16(3): 329-337.

Xie X, Guo X, Liu G, et al. Implicit block diagonal low-rank representation [J]. IEEE Transactions on Image Processing, 2017, 27(1): 477-489.

Xie Z, Li M. Semi-supervised AUC optimization without guessing labels of unlabeled data [C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,. 2018a: 4310-4317.

Xie Z, Li M. Cutting the software building efforts in continuous integration by semi-supervised online auc optimization [C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. 2018b: 2875-2881.

Xin B, Wang Y, Gao W, et al. Data-dependent sparsity for subspace clustering [C]//UAI. 2017.

Xu D, Ouyang W, Wang X, et al. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing [C]//CVPR. 2018: 675-684.

Xu L, Huang A, Chen J, et al. Exploiting task-feature co-clusters in multi-task learning [C]//AAAI. 2015: 1931-1937.

Xue Y, Liao X, Carin L, et al. Multi-task learning for classification with dirichlet process priors [J]. Journal of Machine Learning Research, 2007, 8(Jan): 35-63.

Yan L, Dodier R H, Mozer M, et al. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic [C]//International Conference on Machine Learning. 2003: 848-855.

Yang B. The extension of the area under the receiver operating characteristic curve to multi-class problems [J]. Isecs International Colloquium on Computing, Communication, Control, and Management, 2009: 463-466.

Yang H, Lu K, Lyu X, et al. Two-way partial auc and its properties [J]. Statistical methods in medical research, 2019, 28(1): 184-195.

Yang Z, Zhang T, Lu J, et al. Optimizing area under the ROC curve via extreme learning machines [J]. Knowl.-Based Syst., 2017, 130: 74-89.

Yang Z, Xu Q, Cao X, et al. From common to special: When multi-attribute learning meets personalized opinions [C]//AAAI. 2018: 515-522.

Yang Z, Xu Q, Zhang W, et al. Split multiplicative multi-view subspace clustering [J]. IEEE Transactions on Image Processing, 2019, 28(10): 5147-5160.

Yao Y, Cao J, Chen H. Robust task grouping with representative tasks for clustered multi-task learning [C]//KDD. 2019: 1408-1417.

Yin Z, Shen Y. On the dimensionality of word embedding [C]//NeurIPS. 2018: 887-898.

Ying Y, Wen L, Lyu S. Stochastic online auc maximization [J]. Advances in Neural Information Processing Systems, 2016: 451-459.

You C, Robinson D P, Vidal R. Scalable sparse subspace clustering by orthogonal matching pursuit [C]//CVPR. 2016: 3918-3927.

Yu S, Tresp V, Yu K. Robust multi-task learning with *t*-processes [C]//ICML. 2007: 1103-1110.

Yu Y, Wang T, Samworth R J. A useful variant of the davis–kahan theorem for statisticians [J]. Biometrika, 2014, 102(2): 315-323.

Zeng J, Lau T T, Lin S, et al. Global convergence of block coordinate descent in deep learning [C]// ICML. 2019.

Zhang X D. The laplacian eigenvalues of graphs: a survey [J]. arXiv preprint arXiv:1111.2897, 2011.

Zhang Y, Yang Q. A survey on multi-task learning [J]. arXiv preprint arXiv:1707.08114, 2017.

Zhao P, Hoi S C, Jin R, et al. Online auc maximization [C]//International Conference on Machine Learning. 2011: 233-240.

Zhong W, Kwok J T. Convex multitask learning with flexible task clusters [C]//ICML. 2012: 483-490.

Zhou J, Chen J, Ye J. Clustered multi-task learning via alternating structure optimization [C]//NIPS. 2011: 702-710.

Zhou Q, Zhao Q. Flexible clustered multi-task learning by learning representative tasks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 266-278.

Zou F, Shen L, Jie Z, et al. A sufficient condition for convergences of adam and rmsprop [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 11127-11135.

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

杨智勇，北京人，中国科学院信息工程研究所博士研究生。

已发表(或正式接受)的学术论文:

第一作者：

1. **Zhiyong Yang**, Qianqian Xu, Xiaochun Cao and Qingming Huang. Task-Feature Collaborative Learning with Application to Personalized Attribute Prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020. (In Press)
2. **Zhiyong Yang**, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao and Qingming Huang. When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC. ICML 2021 (Long talk)
3. **Zhiyong Yang**, Qianqian Xu, Yangbangyan Jiang, Xiaochun Cao and Qingming Huang. Generalized Block-Diagonal Structure Pursuit: Learning Soft Latent Task Assignment against Negative Transfer. Annual Conference on Neural Information Processing Systems (NeurIPS), 5846–5857, 2019. (Poster)
4. **Zhiyong Yang**, Qianqian Xu, Xiaochun Cao and Qingming Huang. Learning Personalized Attribute Preference via Multi-task AUC Optimization. AAAI Conference on Artificial Intelligence (AAAI), 5660–5667, 2019. (Oral)
5. **Zhiyong Yang**, Qianqian Xu, Weigang Zhang, Xiaochun Cao and Qingming Huang. Split Multiplicative Multi-view Subspace Clustering. IEEE Transactions on Image Processing (TIP), 28(10): 5147–5160, May 2019. (Regular paper)
6. **Zhiyong Yang**, Qianqian Xu, Xiaochun Cao and Qingming Huang. From Common to Special: When Multi-Attribute Learning Meets Personalized Opinions. AAAI Conference on Artificial Intelligence (AAAI), 515–522, 2018. (Poster)

其他作者：

1. Qianqian Xu, **Zhiyong Yang**, Yangbangyan Jiang, Xiaochun Cao, Yuan Yao and

Qingming Huang. Not All Samples are Trustworthy: Towards Deep Robust SVP Prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021. (Early Access)

2. Qianqian Xu, **Zhiyong Yang**, Zuyao Chen, Yangbangyan Jiang, Xiaochun Cao, Yuan Yao and Qingming Huang. Deep Partial Rank Aggregation for Personalized Attributes. AAAI 2021 (Accepted).
3. Yangbangyan Jiang, Qianqian Xu, Ke Ma, **Zhiyong Yang**, Xiaochun Cao and Qingming Huang. What to Select: Pursuing Consistent Motion Segmentation from Multiple Geometric Models. AAAI 2021 (Accepted).
4. Zongsheng Cao, Qianqian Xu, **Zhiyong Yang**, Xiaochun Cao and Qingming Huang. Dual Quaternion Knowledge Graph Embeddings. AAAI 2021 (Accepted).
5. Qianqian Xu, Jiechao Xiong, **Zhiyong Yang**, Xiaochun Cao, Qingming Huang and Yuan Yao. Who Likes What? – SplitLBI in Exploring Preferential Diversity of Ratings. AAAI Conference on Artificial Intelligence (AAAI), 262–269, 2020. (Oral)
6. Tianwei Cao, Qianqian Xu, **Zhiyong Yang** and Qingming Huang. Task-distribution-aware Meta-learning for Cold-start CTR Prediction. ACM Conference on Multimedia (ACM MM), 3514–3522, 2020.
7. Qianqian Xu, Xinwei Sun, **Zhiyong Yang**, Xiaochun Cao, Qingming Huang and Yuan Yao. iSplit LBI: Individualized Partial Ranking with Ties via Split LBI. Annual Conference on Neural Information Processing Systems (NeurIPS), 3896–3906, 2019. (Poster)
8. Qianqian Xu, **Zhiyong Yang**, Yangbangyan Jiang, Xiaochun Cao, Qingming Huang and Yuan Yao. Deep Robust Subjective Visual Property Prediction in Crowdsourcing. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 8993–9001, 2019. (Poster)
9. Yangbangyan Jiang, Qianqian Xu, **Zhiyong Yang**, Xiaochun Cao and Qingming Huang. DM2C: Deep Mixed-Modal Clustering. Annual Conference on Neural Information Processing Systems (NeurIPS), 5880–5890, 2019. (Spotlight)

10. Ke Ma, Qianqian Xu, **Zhiyong Yang** and Xiaochun Cao. Less but Better: Generalization Enhancement of Ordinal Embedding via Distributional Margin. AAAI Conference on Artificial Intelligence (AAAI), 2978–2985, 2019. (Poster)
11. Shilong Bao, Qianqian Xu, Ke Ma, **Zhiyong Yang**, Xiaochun Cao and Qingming Huang. Collaborative Preference Embedding against Sparse Labels. ACM Conference on Multimedia (ACM MM), 2079–2087, 2019.
12. Yangbangyan Jiang, Qianqian Xu, **Zhiyong Yang**, Xiaochun Cao and Qingming Huang. Duet Robust Deep Subspace Clustering. ACM Conference on Multimedia (ACM MM), 1596–1604, 2019. (Spotlight)
13. Qianqian Xu, Jiechao Xiong, Xinwei Sun, **Zhiyong Yang**, Xiaochun Cao, Qingming Huang and Yuan Yao. A Margin-based MLE for Crowdsourced Partial Ranking. ACM Conference on Multimedia (ACM MM), 591–599, 2018. (Full paper)
14. Yangbangyan Jiang, **Zhiyong Yang**, Qianqian Xu, Xiaochun Cao and Qingming Huang. When to Learn What: Deep Cognitive Subspace Clustering. ACM Conference on Multimedia (ACM MM), 718–726, 2018.

参与项目:

1. 2020/01-2023/12 国家自然科学基金面上基金，61976202，基于网络众包的主观视觉属性偏序算法研究
2. 2018/01-2020/12 北京市自然科学基金面上基金，4182079，基于几何拓扑的网络众包数据分析
3. 2019/12-2023/12 科技创新2030-“新一代人工智能”重大项目面向跨媒体内容管理的智能分析与推理 2018AAA0102000

获奖情况:

1. 百度全球20强（面向全球华人），2019
2. 中科院院长特别奖（中科院及中科大共62人），2020
3. NeurIPS 2020 Top-10% 审稿人， 2020
4. 中科院信息工程研究所所长特别奖，2019

5. 中科院信息工程研究所所长特别奖，2018
6. 博士研究研究生国家奖学金，2019
7. 博士研究研究生国家奖学金，2018

学术服务:

1. IJCAI 2021 Senior Program Committee Member (SPC)
2. ICML 2021 Expert Reviewer (ER)
3. NeurIPS 18/19/20, ICML 19/20, AAAI 20/21 及IEEE Transactions on Image Processing 审稿人

致　谢

四年的普博读生涯随着学位论文的完成也接近尾声，在此由衷感谢攻读博士学位期间所有关心、帮助、支持我的老师、同学和家人们!

首先，我要衷心感谢我的导师黄庆明教授。黄老师敏锐的思维、严谨的治学态度、渊博的学识、诚挚谦虚的品格和宽厚善良的处世方式，永远值得我学习和效仿。再次向黄老师所有的付出表示衷心地感谢!

衷心感谢课题小组长许倩倩老师对我的谆谆教诲。记不清有多少次在许老师的耐心指导下做实验，改论文。许老师高标准的研究水平、丰富的研究经验、忘我的工作精神一直是我钦佩和学习的榜样。感谢实验室的操晓春、王树徽、李亮、李国荣、张华、任文琦老师，与各位老师的学习和交流，使我终生受益。他们的鼎力帮助和热情关心，使我终生难忘。

感谢我的师兄马坷博士、吴哲博士、齐元凯博士、独大为博士，他们从一个先行者的角度给了我无数的建议和关心，所有这些都使我有幸绕过了很多研究中的急流险滩，为我能顺利完成博士研究提供了莫大的帮助。感谢实验室与我朝夕相处多年的师妹姜阳邦彦、郝前秀，师弟包世龙、王子泰、曹天伟及其实验室其他同学在科研和生活中给予我的帮助，与他们在一起的每个日子都将成为我生命里美好的回忆。

最后，向百忙中抽出宝贵时间评审本论文的专家和学者致以最诚挚的谢意。