基于稳定替代损失的可泛化平均精度优化

温佩松[1),2)] 许倩倩[1)] 杨智勇[2)] 黄庆明[1),2)]

[1)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[2)](中国科学院大学计算机科学与技术学院 北京 100190)

摘 要 平均精度(Average Precision,AP)由于其对排序性能的全面度量,已经成为多种计算机视觉任务中广泛使用的验证指标,包括长尾分类、图像检索和目标检测等。为缩小训练目标与验证指标之间的差距,近年研究提出了 AP 指标的直接优化算法。然而,受限于 AP 风险的不可分解性,大多数现有的 AP 优化方法基于不稳定的替代损失,即更改一个样本可能导致损失估计大幅波动。受该特性影响,期望风险与经验风险差距可能受到少数异常样本的影响,致使模型泛化能力欠佳。鉴于此,本文旨在探索一种适用于随机 AP 优化的可泛化算法。为了克服不稳定问题,本文首先基于 AP 风险的紧上界推导出一种加权成对形式的替代目标,使其具有良好的稳定特性。在理论方面,本文对所提出的替代目标展开泛化分析,证明最小化替代经验风险可有效优化原目标的期望风险。在此基础上,为有效最小化替代风险,本文设计了一种具有可证明收敛性的随机优化算法。在实践方面,通过 3 种任务、7 个基准数据集上的全面实验验证了所提出框架的有效性和理论结果的可靠性。

关键词 平均精度优化;排序学习;泛化性理论;图像检索

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2025.02094

Generalizable Stochastic Average Precision Learning via Stable Surrogate Loss

WEN Pei-Song[1),2)] XU Qian-Qian[1)] YANG Zhi-Yong[2)] HUANG Qing-Ming[1),2)]

[1)](Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

[2)](School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190)

Abstract Benefiting from its comprehensive measures of ranking performance, Average Precision (AP) has become a widely used evaluation metric in computer vision tasks, such as long-tailed classification, image retrieval, and object detection. Recent methods propose direct AP optimization algorithms to narrow the gap between training objectives and evaluation metrics. However, limited by the non-decomposable AP risks, the majority of existing AP optimization methods suffer from unstable losses, i.e., changing one sample might lead to large jitters in loss estimations. Such a property reduces the generalization since the expected/empirical risk gap might be affected by a few abnormal samples. In this paper, we aim to explore a generalizable algorithm for stochastic AP optimization. To overcome the unstable issue, we first derive a surrogate objective with a weighted pairwise formulation from a tight upper bound of the AP risk. Theoretically, we provide a generalization analysis for the proposed surrogate objective. Then a stochastic optimization algorithm is designed for our surrogate objective with a provable convergence. Practically, comprehensive experiments over 7 benchmarks of 3 tasks speak to the effec-

收稿日期:2024-08-04;在线发布日期:2025-04-14。本课题研究由科技创新 2030-"新一代人工智能"重大项目(2018AAA0102000)、国家自然科学基金项目(62441232,62236008,U21B2038,U23B2051,62122075)、中国科学院青年促进会优秀会员项目、中国科学院战略性先导科技专项(XDB0680201)。**温佩松**,博士,特别研究助理,中国计算机学会(CCF)会员,主要研究领域为机器学习、计算机视觉。E-mail:wenpeisong20z@ict.ac.cn。**许倩倩(通信作者)**,博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为统计机器学习及其在多媒体领域的应用。E-mail:xuqianqian@ict.ac.cn。**杨智勇**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为机器学习理论。**黄庆明(通信作者)**,博士,教授,中国计算机学会(CCF)会士,主要研究领域为多媒体计算、图像处理、计算机视觉与模式识别。E-mail:qmhuang@ucas.ac.cn。

tiveness of the proposed framework and the soundness of theoretical results.

Keywords　average precision optimization; learning to rank; generalization theory; image retrieval

1 引　言

作为精度-召回率曲线下面积(Area Under the Precision-Recall Curve,AUPRC)的无偏估计,平均精度(Average Precision,AP)度量了精度和召回率的综合性能,被认为是一个更全面的指标,尤其对于高度不平衡的数据集[1]。凭借这一吸引人的特性,AP 通常被视为图像检索[2-3]、目标检测[4-5]和分类[6-7]等各种视觉任务中的标准指标。

AP 的广泛应用在过去二十年引发了对 AP 直接优化方法的研究热潮。早期研究集中于全批量方法[8-10]。然而,这些方法存在计算复杂度高的局限,难以扩展到现代深度学习框架中。近期深度学习的快速发展使得对随机 AP 优化的关注日益增加[2-3,5,11-13]。例如,针对 AP 指标不可微问题,采用 Sigmoid 函数[2,12]或指数函数[14]等可微替代损失替换不可微的 0-1 损失。从 PR 曲线的定义出发,其他方法[3,14-15]使用直方图分箱技术[16]来近似精确率和召回率。Burges 等人[17-18]则采用零阶有限差分近似不可微函数的梯度。上述方法均较为直观,并且提高了某些计算机视觉任务中的 AP 性能,但受限于不稳定的损失函数,大多数方法仍缺乏适当的泛化保证。鉴于此,本文聚焦于下列问题:

如何设计一个具有可证明泛化保证的 AP 优化算法?

以往关于信息检索的工作[19-21]已经在双层问题中研究了 AP 的泛化,即数据集包含多个查询,每个查询包含多个二分类候选样本(是否相关)。当前研究表明,该问题设定下,仅当查询数量足够大时,才具有较小的泛化误差上界。然而,上述理论结果无法解释单列表问题的泛化性,即数据集仅包含少量或单个查询,例如目标检测或长尾分类问题。为弥补这一空白,本文提供了一个初步尝试,研究单列表级别的 AP 优化算法泛化性。一方面,这要求可证明的泛化界,即替代经验风险与原期望风险的差距可控;另一方面,要求适当的收敛速度,即替代经验风险可被充分优化。达成上述目标的主要挑战包含两方面:(1)理论层面,AP 指标无法分解为仅包含单个样本的独立项,使得标准的泛化分析工具[22]不适用。此外,常用替代损失与原始 AP 风险之间的关系尚不明确,进一步增加理论保障难度;(2)方面层面,AP 的经验替代损失涉及所有样本的排序,这需要通过低效的全样本扫描以获得无偏估计。

为了解决问题(1),本文从 AP 损失的一个性质出发:它可以被看作是一种排序加权的成对损失。基于这一事实,第 4 节中推导出了一个具有排序加权成对形式的 AP 损失的上界。基于此,通过权重估计,AP 损失被分解成多个独立分量组成的稳定损失,为后续在第 5 节中得到可证明的泛化界奠定基础。

在此基础上,第 6 节重点解决挑战(2),即如何高效地优化本文提出的替代损失。首先,本文将待最小化的替代目标重构为双层优化问题。随后,在第 6.1 小节中提出了一个随机优化框架来联合优化这两个层次。第 6.2 小节中从理论上证明了本文提出的算法能够在 $O(1/\epsilon^4)$ 样本复杂度下使优化目标达到一个ϵ-驻点。

综上所述,本文的主要贡献如下:

(1)稳定的 AP 替代优化目标。本文从 AP 风险的上界推导出一个稳定的 AP 替代损失,从而获得可证明的泛化保障。基于此,本文提出了单列表 AP 优化泛化界,克服了传统方法对查询数充分大的要求。

(2)收敛性良好的优化算法。针对所提出替代目标的双层优化特性,本文设计了无偏随机优化算法,并证明该算法具有良好的收敛率。

(3)多任务应用验证。在实践方面,本文在图像检索、目标检测、长尾分类等任务的 7 个基准数据集上进行了全面的实验研究,展示了所提出框架的广泛应用潜力。

2 研究现状

2.1 AUROC 优化方法

除了平均精度(AP)之外,在排序学习和长尾问题中另一个常用指标是 ROC 曲线下面积(AUROC),它具有对长尾数据分布不敏感的特性,因此被视为长尾问题的标准指标之一。早期关于 AU-

ROC 优化的研究[23-24]聚焦于在全批量学习,计算量较大。为提升效率,研究人员提出多项式近似方法[25]和基于集成的方法[26]求解 AUROC 优化问题。在过去几十年中,在线和随机方法[27-31]相继被提出,将 AUROC 优化扩展到深度学习框架下的大规模数据集中。最近,Ying 等人[32]将随机 AUROC 优化问题重构为随机鞍点问题,该形式在相关问题中被进一步发展利用[28-29]。

2.2 AP/AUPRC 优化方法

尽管 AUROC 在机器学习中广泛用作测试指标,但当数据分布严重不平衡时,它会倾向于过于乐观[1]。因此,文献[22-33]使用精确率-召回率(Precision-Recall,PR)曲线代替 ROC 曲线来衡量模型在严重不平衡数据集上的性能,导出名为 PR 曲线下面积(Area Under the Curve, AUPRC)的测试指标。对于有限数据集,AUPRC 可采用平均精度(Average Precision,AP)作为无偏估计[34-36],因此针对 AUPRC 指标的优化方法通常被称为 AP 优化。由于 AP 和 AUROC 的不一致性[1],如何直接优化 AP 在过去几十年中引起了研究人员的关注。针对 AP 指标既不可微分也不可分解的特点,一些早期工作利用离散优化方法来解决这个问题,例如马尔可夫随机场模型[9]、随机搜索[8]、动态规划[37]和误差驱动方法[17]。作为早期尝试,Mohapatra 等人[10]通过最小化 AP 的正则化凸上界来探索在 SVM 中优化 AP。然而,由于高复杂度和对模型原型的强假设,这些方法难以到深度学习。因此,更多方法转向在深度学习框架中优化 AP,特别是在计算机视觉领域。

在深度学习中的 AP 优化方法大致可以分为三条技术路线:(1)平滑近似,(2)误差驱动方法,(3)黑盒优化。技术路线(1)的关键思想是用平滑的替代目标来近似不可微的目标,以便实现基于梯度的优化,例如,用 Sigmoid 函数[2,12]或指数函数[14]替换 0-1 损失。从 PR 曲线的定义出发,一些其他方法[3,14-15]使用直方图分箱技术[16]来近似精确率和召回率。这些方法比较直观,并且提高了某些计算机视觉任务中的 AP 性能,但仍缺乏泛化的理论保证。相比之下,Burges 等人[17-18]提出技术路线(2),旨在避免直接从不可微目标计算梯度。具体来说,根据链式法则,通过用差分替换关于得分函数的微分,可得关于参数的梯度。这一思想后来被扩展用于解决目标检测中的长尾分布问题[4-5,38]。这些方法避免了人工设计的替代函数,但仅在线性模型上有较弱的收敛性保证[5]。而对于技术路线(3),黑盒优化[39-41]旨在解决一般的组合优化问题,因此可以用于排序优化[42],但由于追求算法通用性,忽略了对排序学习问题的一些基本性质,性能仍有待提升。

本文在平滑近似路线的基础上更进一步。为填补 AP 优化的理论空白,提出了一个 AP 指标的稳定上界作为替代目标,并为所提出的损失函数提供了一个泛化界。此外,针对目标函数的双层特性提出了高效的一阶优化算法,并提供了相应的收敛性理论保障。

2.3 排序学习的泛化性理论

泛化性能,即模型在未知测试数据上的性能,是机器学习的一个基本问题。作为泛化分析标准框架的关键组成部分,Rademacher 复杂度[43-45]被提出用于衡量模型复杂度。不幸的是,原始的 Rademacher 复杂度要求待分析经验风险可被表示为独立项之和,而成对目标函数和列表目标函数并不满足该假设。因此,研究人员进一步开发了更多针对成对或列表目标函数的技术。

以 AUROC 为典范,首先介绍关于成对目标函数泛化界的研究。为了将 AUROC 风险分解为一系列标准逐项风险之和,文献[1,46]利用图着色技术将相互依赖的项划分为若干组,使得同一组中的各项独立。基于此,每组均分别适用标准框架。此外,为解决该问题,文献[46,30]为 AUROC 风险提出了新的 Rademacher 复杂度变体。

除了成对风险之外,列表风险的泛化界在排序学习中也非常重要。例如,常用的指标如平均精度和归一化折损累计增益(Normalized Discounted Cumulative Gain,NDCG)均表示为相互依赖的排序列表项之和,即每一项均涉及排序列表中所有样本,难以解耦。信息检索领域(Information Retrieval, IR)的文献[17,47]研究了双层设定下的列表泛化界,即关于查询数量(查询层次)和每个查询的文档列表长度(列表层次)的泛化。在查询是独立同分布采样的假设下,Lan 等人[47]提供了典型排序方法[19-21]的泛化上界。Chen 等人[48]弱化了查询是条件独立同分布采样的假设,并且提供了一个 $O\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)$ 阶的泛化上界,其中 m 和 n 分别是查询数量和列表长度。这个结论要求同时满足 $m \to \infty$ 和 $n \to \infty$,才能使泛化界一致收敛。Tewari 和 Chaudhuri[49]进一步探索了 $O\left(\frac{1}{\sqrt{m}}\right)$ 阶的列表泛化上界,可随查询

数量增加时收敛,不依赖列表长度。

为了确保 AP 在目标检测、长尾分类等查询有限场景中的泛化性能,本文进一步研究了单列表层次上的 AP 泛化界。该问题更具挑战性,因为目标函数中所有项均相互依赖,使得部分依赖风险技术(例如图着色技术)不可行。为了解决这个问题,本文提出一种排序加权的成对函数作为替代目标。基于成对 Rademacher 复杂度[30],本文提出一种加权成对 Rademacher 复杂度,进而推导出 $O\left(\frac{1}{\sqrt{n}}\right)$ 阶的泛化界,克服现有方法对查询数依赖。

3　平均精度的性质

3.1　预备知识

给定一个二分类数据集 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}\sim\mathcal{D}$,其中 $\mathcal{D}$ 为数据分布空间,$\boldsymbol{x}\in\mathbb{R}^d$ 为 d 维特征,$y\in\{0,1\}$ 为二分类标签。$\mathcal{S}$ 的特征可以分为正子集 $\mathcal{S}^+=\{\boldsymbol{x}_i^+\}_{i=1}^{n^+}$ 和负子集 $\mathcal{S}^-=\{\boldsymbol{x}_i^-\}_{i=1}^{n^-}$,分别采样自两个不同的分布:$\boldsymbol{x}_i^+\overset{i.i.d.}{\sim}\mathcal{P}$:$\mathbb{P}[\boldsymbol{x}^+\mid y=1]$,$\boldsymbol{x}_i^-\overset{i.i.d.}{\sim}\mathcal{N}$:$\mathbb{P}[\boldsymbol{x}^-\mid y=0]$,其中 n^+,n^- 是两个子集的大小,满足 $n=n^++n^-$。本文目的为学习一个评分函数 $f_{\boldsymbol{\theta}}\in\mathcal{F}$:$\mathbb{R}^d\mapsto\mathbb{R}$,其中 $\boldsymbol{\theta}$ 是从假设集 $\mathcal{F}$ 中选择的参数,使得对于任意正负样本对 (x_i^+,x_j^-),尽可能满足 $f_{\boldsymbol{\theta}}(x_i^+)>f_{\boldsymbol{\theta}}(x_j^-)$。

为了便于表达,下文在无歧义的情况下简记 $\mathcal{S}=\mathcal{S}^+\cup\mathcal{S}^-$,$f=f_{\boldsymbol{\theta}}$,并附加如下符号表示:记 $\rho=n^-/n^+$,$\delta_{ij}^{+}=f(\boldsymbol{x}_i^+)-f(\boldsymbol{x}_j^+)$,$\delta_{ij}^{+-}=f(\boldsymbol{x}_i^+)-f(\boldsymbol{x}_j^-)$;设 $\ell_{0,1}$ 为 0-1 损失函数,即当 $x\leqslant 0$ 时 $\ell_{0,1}(x)=1$,否则 $\ell_{0,1}(x)=0$;给定一个有限数据集 $\mathcal{S}$,记实例 x 在 $\mathcal{S}$ 上的排名为 $\mathrm{rank}(x,\mathcal{S};f)=\sum_{x_j\in\mathcal{S}}\ell_{0,1}(f(x_j)-f(x))$。

基于上述符号,AUPRC[50] 定义如下,其中 t 是取值于 $(-\infty,\infty)$ 的决策阈值:

$$\mathrm{AUPRC}(f)=\int_0^1\underbrace{\mathbb{P}[y=1\mid f(x)>t]}_{\text{精度}}\mathrm{d}\underbrace{\mathbb{P}[f(x)>t\mid y=1]}_{\text{召回率}},$$

上述定义基于连续的样本分布,而给定有限数据集,通常采用平均精度(AP)作为 AUPRC 的一种无偏经验估计[34],形式如下:

$$\widehat{AP}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}\left(\mathrm{rank}(x_i^+,\mathcal{S}^+;f)/\mathrm{rank}(x_i^+,\mathcal{S};f)\right)。$$

显然,最大化评分函数 f 的期望 AP 等价于最小化以下风险:

$$(OP0)\ \min_{f\in\mathcal{F}}\mathfrak{R}^{\ell_{0,1}}(f)=\mathbb{E}_{\mathcal{S}}[\hat{\mathfrak{R}}_{\mathcal{S}}^{\ell_{0,1}}(f)],$$

$$\hat{\mathfrak{R}}_{\mathcal{S}}^{\ell_{0,1}}(f)=1-\widehat{AP}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}\left(\frac{\sum_{j=1}^{n^-}\ell_{0,1}(\delta_{ij}^{+-})}{\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^{+})}\right),$$

其中,$\tilde{h}(x)=x/(1+x)$ 为单调递增函数。详细推导见附录 A。

3.2　排序指标的不一致性

由于 AP 指标的形式较为复杂,一种常用方法是优化其他排序指标[31,51-52],例如接收者操作特征曲线下面积(Area Under the Receiver Operating Characteristic,AUROC)以及前 K 准确率(Precision at K,Prec@K),形式如下:

$$\mathrm{AUROC}(f)=\frac{1}{n^+n^-}\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}1[f(\boldsymbol{x}_i^+)>f(\boldsymbol{x}_j^-)],$$

$$\mathrm{Prec@K}(f)=\frac{1}{K}\sum_{i=1}^{n^+}1[\mathrm{rank}(\boldsymbol{x}_i^+,\mathcal{S};f)\leqslant K]。$$

然而,本文将展示这些指标与 AP 存在不一致现象,即在上述指标上性能好的模型,可能 AP 指标反而较低。如图 1 所示,考虑三个评分函数。在 AP 指标下,偏好顺序为 $f_1\succ f_2\succ f_3$,但是在 AUROC 指标下恰好相反。Prec@K 则对模型性能的敏感度较低,因为它只计算前 K 个列表中的正样本数量,而忽略了内部的相对排序。这种不一致性揭示了直接优化 AP 的必要性。

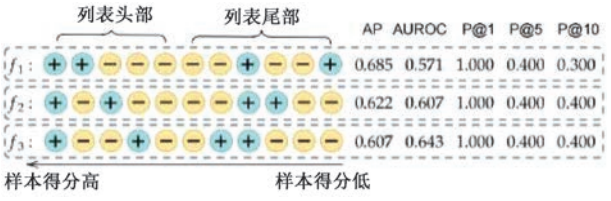


图 1　AP 指标与其他排序指标(AUROC、Prec@K)的不一致性。AP 指标聚焦于排序列表头部性能。

上述例子引发的一个问题是:既然这些排序指标不一致,AP 应该应用于哪些场景?与 f_2 和 f_3 相比,f_1 在头部表现更好,而在尾部表现相对较差。就 AP 指标的定义而言,排序靠前的错分样本对权重更大,因此 AP 侧重于头部,并且鼓励提升高分正样本的表现,故 AP 更适合于检索、目标检测等对得分高度敏感的场景。

3.3　AP 损失的稳定性

一般而言,AP 优化算法旨在最小化其原始期望风险 $\mathfrak{R}^{\ell_{0,1}}(f)$,但实际中只能优化定义在有限数

据集、可微替代损失上的替代经验风险，导致泛化误差。聚焦于该问题，本文指出分析泛化误差与损失函数的稳定性密切相关。直觉上，泛化误差受训练数据的质量主导，异常样本、标注噪声等可能导致替代损失波动，而不稳定损失波动更明显，导致优化更新偏离期望方向，进而弱化模型在测试数据上的泛化能力。因此，稳定 AP 替代损失对模型在 AP 指标上的泛化性至关重要。

为定量研究上述问题，首先，本文利用有界差分性质(Bounded Difference Property，BDP)来度量损失函数稳定性，具体定义如下：

定义 1. 有界差分性质. 对于任意两个仅相差一个样本的数据集 $\mathcal{S}$ 和 $\mathcal{S}'$，如果存在一个非负常数 c，目标函数 $\mathcal{R}_{\mathcal{S}}^{\ell}(f)$：$\mathcal{X}^n \mapsto \mathbb{R}$ 满足以下条件，则称其具备有界差分性质：

$$\sup_{f,\mathcal{S},\mathcal{S}'} \left| \mathcal{R}_{\mathcal{S}}^{\ell}(f) - \mathcal{R}_{\mathcal{S}'}^{\ell}(f) \right| \leqslant c。$$

考虑经典的逐实例损失，例如交叉熵损失：$\mathcal{R}_{\mathcal{S}}^{\ell}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_{ce}(f(\boldsymbol{x}_i), y_i)$。BDP 常数 c 是 $O(1/n)$，因为替换一个样本后仅影响一项。直观上，在样本充足的情况下，目标函数对训练样本的微小变化不敏感，泛化误差较小(对于预期的实例风险)。

对于 AP 损失，其包含的 0-1 损失不可微分，难以通过端到端方法直接最小化 $\mathcal{R}_{\mathcal{S}}^{\ell_{0,1}}(f)$。初步观察，可以采用平滑替代函数 ℓ 替换 $\ell_{0,1}$[2,12,14]，从而得到可微目标函数：

$$\mathcal{R}_{\mathcal{S}}^{\ell}(f) = \frac{1}{n^+}\sum_{i=1}^{n^+} \tilde{h}\left(\frac{\sum_{j=1}^{n^-} \ell(\delta_{ij}^{+-})}{\sum_{j=1}^{n^+} \ell(\delta_{ij}^{+})} \right)。$$

然而，即使只替换一个样本，上述目标也可能发生较大变化，换言之，BDP 常数为 $O(1)$，不随样本数增加而递减。这是由于其不可分解形式：求和式中的每一项都涉及所有样本，因此改变一个样本会影响所有项。这种特性是通向可泛化 AP 损失的主要瓶颈，本文旨在克服该不足之处。

4 稳定 AP 替代目标

为了确保可证明的泛化性能，即通过最小化替代经验风险 $\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f)$ 可充分优化原期望风险 $\mathcal{R}^{\ell_{0,1}}(f)$，下文聚焦于这两项之间的数量关系。结合第 3.3 小节中提到的稳定性问题，AP 替代目标基于以下两个原则展开：

(1)目标函数应满足良好稳定性，即 BDP 常数随训练样本量递减，进而可通过增加样本量缩小替代经验风险 $\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f)$ 和替代期望风险 $\mathcal{R}^{\ell}(f)$ 之间的差距；

(2)替代期望风险 $\mathcal{R}^{\ell}(f)$ 和原期望风险 $\mathcal{R}^{\ell_{0,1}}(f)$ 之间的差距可控，以确保优化目标和评估指标的一致性。

基于以上原则，文本围绕稳定 AP 替代目标设计及其优化算法展开，并应用于多种下游任务，研究整体框架如图 2 所示。

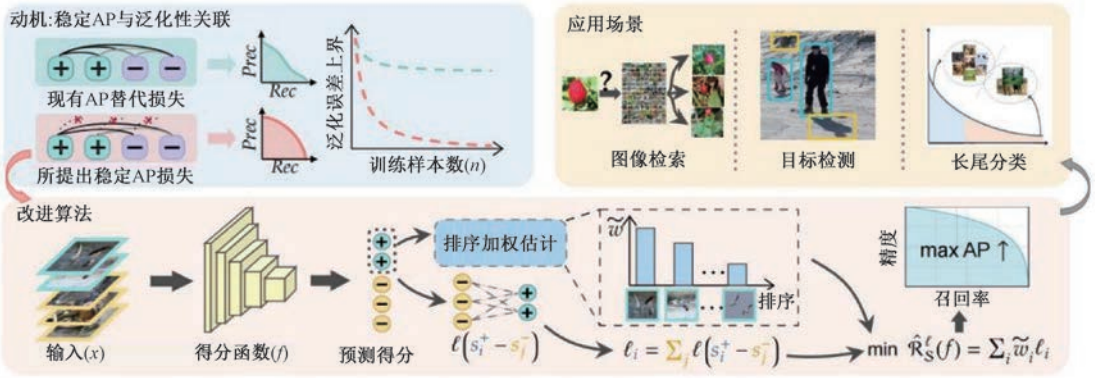


图 2 研究内容总览(本文发现 AP 损失稳定性与泛化性密切相关，随着训练样本数增加，稳定 AP 损失的泛化误差上界收敛到零。鉴于此，提出通过成对样本损失与排序加权来估计稳定 AP 损失，并基于此提出相应双层优化算法。所提出算法被应用于多种视觉任务，包括图像检索、目标检测和长尾分类。)

4.1 AP 损失的稳定上界

为寻找一个稳定的 AP 替代损失，本文从第 3.2 小节中提到的一个 AP 指标性质出发，即它通过控制错分样本对的权重聚焦于排序列表头部性

能。直观而言,为了最小化原始经验风险 $\hat{\mathcal{R}}_S^{\ell_{0,1}}(f)$,应尽可能使正样本得分高于负样本得分,而交换一对正样本的得分并不会改变 AP。因此,正样本间的排序 $\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^+)$ 可视为权重,而非优化目标。鉴于此,本文通过如下技术路线推导稳定 AP 替代目标:首先,重构 AP 指标为带权成对损失形式,将 AP 指标中样本间的复杂耦合关系限制在损失权重上;随后,通过估计权重消除耦合关系,从而确保 AP 替代目标的稳定性。

具体地,记 $r_i=\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^+)/n^+$ 为正样本的归一化排序,则 $\hat{\mathcal{R}}_S^{\ell_{0,1}}(f)$ 可重构为

$$\hat{\mathcal{R}}_S^{\ell_{0,1}}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}\left(\frac{\rho}{n^-}\sum_{j=1}^{n^-}\underbrace{1/r_i}_{\text{损失权重}}\cdot\underbrace{\ell_{0,1}(\delta_{ij}^{+-})}_{\text{成对损失}}\right)\tag{1}$$

以平滑的代理损失 ℓ 替换式(1)中的 0−1 损失,可得可微目标函数。记 $\ell_i=\frac{1}{n^-}\sum_{j=1}^{n^-}\ell(\delta_{ij}^{+-})$,通过选择满足 $\ell\geqslant\ell_{0,1}$ 的代理损失,可得

$$\hat{\mathcal{R}}_S^{\ell_{0,1}}(f)\leqslant\tilde{\mathcal{R}}_S^{\ell}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}(\rho/r_i\cdot\ell_i)\tag{2}$$

上述结论表明,通过最小化 $\tilde{\mathcal{R}}_S^{\ell}$ 可有效优化原经验风险 $\hat{\mathcal{R}}_S^{\ell_{0,1}}$。然而,$\tilde{\mathcal{R}}_S^{\ell}$ 仍然不是合适的替代目标函数,因为非线性函数 $\tilde{h}$ 分离了对正样本的求和与对负样本的求和,导致较弱的 BDP。注意到 $\tilde{h}$ 是凹函数,一个满足原则(1)的自然替代方法是以下上界:

$$\tilde{U}_S^{\ell}(f)=\tilde{h}\left(\frac{1}{n^+}\sum_{i=1}^{n^+}\rho/r_i\cdot\ell_i\right)\geqslant\tilde{\mathcal{R}}_S^{\ell}(f)\tag{3}$$

式(3)可根据 Jensen 不等式证明。不幸的是,上述不等式放缩过度,相应的损失上界 $\tilde{U}_S^{\ell}(f)$ 过度偏离原损失,进而违背了设计原则(2)(命题完整形式见附录 A)。

为了解决这个问题,本文进一步推导出 $\tilde{\mathcal{R}}_S^{\ell}(f)$ 的紧致上界,如下述定理所示:

定理 1. 设 $\ell:\mathbb{R}\mapsto\mathbb{R}^+$ 为光滑函数,使得对于任意 $x\in\mathbb{R}$ 及 $\ell\in[0,B_\ell]$,都有 $\ell_{0,1}(x)\leqslant\ell(x)$。记 $\rho=n^-/n^+$,$\tilde{h}(x)=\frac{x}{1+x}$,$h(x)=(\epsilon^2+\tilde{h}(x))^{-1/2}$,$\epsilon>0$,以及 $q=\frac{\pi}{2}+\frac{2(1+\rho B_\ell)}{1+2\rho B_\ell}$,如下不等式成立:

$$\hat{\mathcal{R}}_S^{\ell_{0,1}}(f)\leqslant\tilde{\mathcal{R}}_S^{\ell}(f)\leqslant\frac{q}{2}\cdot\hat{\mathcal{R}}_S^{\ell}(f,w)\triangleq\frac{q}{2}\cdot\hat{\mathcal{R}}_S^{\ell}(f,w)$$

,其中,

$$\hat{\mathcal{R}}_S^{\ell}(f,w)=h\left(\frac{\rho}{n^+}\sum_{i=1}^{n^+}w(r_i)\ell_i\right),$$

$$w(r_i)=\begin{cases}\frac{1}{2}(r_i-r_i^{\,2})^{-1/2},0<r_i\leqslant\frac{1}{2},\\(1+\rho B_\ell)/(r_i+\rho B_\ell),\frac{1}{2}<r_i\leqslant 1。\end{cases}$$

上述定理表明,$\hat{\mathcal{R}}_S^{\ell}(f,w)$ 是一个更合适的优化目标,原因如下:首先,$\tilde{h}$ 被一个外部嵌套函数 h 替代,满足原则(1);其次,它是 $\hat{\mathcal{R}}_S^{\ell_{0,1}}(f)$ 的一个紧上界(更多讨论见第 5 节),满足原则(2)。

4.2　目标函数改进

面向广泛应用的深度模型随机优化框架,上述优化目标 $\hat{\mathcal{R}}_S^{\ell}(f,w)$ 仍存在一定局限性,具体体现在其样本排序 $r_i=\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^+)/n^+$ 的计算复杂度高和加权函数 w 的数值稳定性差。因此,本小节将讨论替换上述两个关键组件。

注意到 $\hat{\mathcal{R}}_S^{\ell}(f,w)$ 包含一个基于排序的项 $r_i=\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^+)/n^+$,需要通过全批量扫描获得。为了避免这种情况,采用正例 $\boldsymbol{x}_i^+$ 的期望形式替换 r_i 的经验形式。此外,阶跃函数 $\ell_{0,1}$ 对正样本得分 $f(\boldsymbol{x}_i^+)$ 的微小变化不敏感,导致微小偏差难以纠正。为增强权重的区分能力,本文改用其平滑版本:

$$r_i^{\ell}=\frac{1}{n^+}\mathbb{E}_{\boldsymbol{x}_j^+\sim\mathcal{P}}[\ell(f(\boldsymbol{x}_i^+)-f(\boldsymbol{x}_j^+))/B_\ell],$$

其中,$\ell_{0,1}$ 被替换为值域为 0 到 1 的平滑函数 ℓ/B_ℓ。根据 Jensen 不等式,如果 ℓ 是凸函数,则有 $r_i^{\ell}\geqslant\tilde{r}_i=\ell(f(\boldsymbol{x}_i^+)-\mu^+)/B_\ell$,其中 $\mu^+=\mathbb{E}_{\boldsymbol{x}_j^+\sim\mathcal{P}}[f(\boldsymbol{x}_j^+)]$ 为正样本平均得分。w 的单调性进一步表明 $w(r_i)\approx w(r_i^{\ell})\leqslant w(\tilde{r}_i)$,从而替换后替代目标仍为 AP 风险的上界。通过将 r_i 替换为 $\tilde{r}_i$,只需有效估计 μ^+,即可在无全批量扫描的情况下获得权重,具体算法将在第 6 节中进一步讨论。

实践中发现,定理 1 中定义的加权函数 w 可能会导致梯度爆炸,因为当 $\tilde{r}_i\to 0$ 时 $w(\tilde{r}_i)\to\infty$。此外,其分段形式破坏了加权的平滑性,导致数值不稳定。因此,考虑替换为如下加权函数:

$$\tilde{w}(x)=((1+a)/(x+a))^t\leqslant B_{\tilde{w}}\tag{4}$$

其中，a 和 t 为可调超参数。当 $a=\rho B_\ell$ 并且 $t=1$ 时，对于 $\frac{1}{2}<x\leqslant 1$ 可得 $\tilde{w}(x)=w(x)$，而对于 $0<x\leqslant\frac{1}{2}$ 可以选择较小的 a 和较大的 t，使得 $\tilde{w}(x)$ 与 $w(x)$ 处在同一量级。与 w 类似，$\tilde{w}$ 单调递减。因此，通过设置合适的超参数，$\tilde{w}$ 可以看作是 w 的平滑近似。

综上所述，AP 优化的经验替代目标定义如下：

$$(OP1)\ \min_{f\in\mathcal{F}}\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})=h\left(\frac{\rho}{n^+}\sum_{i=1}^{n^+}\tilde{w}(\tilde{r}_i)\,\ell_i\right)。$$

引理 1. $\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})$ 满足有界差分性质并且 BDP 常数为 $O(1/n)$.

上述引理表明，经过一系列近似后，所得 AP 替代目标仍保持良好稳定性。接下来，本文将表明该替代目标也满足设计原则(2)，从而推导出可证明的泛化界。

5 AP 替代目标泛化性分析

在第 4 节中，本文提供了一个替代目标函数 $\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})$，可视为 AP 经验风险上界的近似，即

$$\hat{\mathcal{R}}^{\ell_{0,1}}_{\mathcal{S}}(f)\leqslant\tilde{\mathcal{R}}^{\ell}_{\mathcal{S}}(f)\leqslant\frac{q}{2}\cdot\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,w)\approx\frac{q}{2}\cdot\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})\tag{5}$$

上述近似带来了一个问题：通过 $\tilde{w}$ 估计基于排序的权重 w 后，能否通过最小化替代经验风险 $\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})$ 显著优化原 AP 期望风险 $\mathcal{R}^{\ell_{0,1}}(f)=\mathbb{E}_{\mathcal{S}}[\hat{\mathcal{R}}^{\ell_{0,1}}_{\mathcal{S}}(f,w)]$？

为了解决这个问题，本节旨在探索所提出目标函数的泛化差异，即证明存在一个正常数 a，使得下式大概率成立：

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell_{0,1}}(f)-a\cdot\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,w)]\leqslant\epsilon(n),$$

其中，当 $n\to\infty$ 时 $\epsilon(n)\to 0$。

上述目标的主要挑战在于成对损失耦合了一对样本。为了度量期望风险和经验风险之间的差距，标准泛化性分析框架利用了 Rademacher 复杂度的对称性[53]。然而，在相互依赖的成对项中缺乏对称性使得标准框架不可行。为了解决这个问题，本文引入成对 Rademacher 复杂度的变体[30]，如下所示：

定义 2. 成对 Rademacher 复杂度. 对于一个二分类数据集 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$，其中 $y_i\in\{-1,+1\}$，以及分数函数 f 的假设空间 $\mathcal{H}$，经验成对 Rademacher 复杂度和期望成对 Rademacher 复杂度分别定义如下：

$$\mathfrak{R}^{p}_{\mathcal{S}}(\mathcal{H})=\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\frac{\sigma_i^+-\sigma_j^-}{2n^+n^-}\cdot\tilde{w}(\tilde{r}_i)\,\ell_{ij}\right|\right],$$

$$\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{H})=\mathbb{E}_{\mathcal{S}\sim\mathcal{D}}[\mathfrak{R}^{p}_{\mathcal{S}}(\mathcal{H})],$$

其中，$\boldsymbol{\sigma}=(\sigma_1^+,\cdots,\sigma_{n^+}^+,\sigma_1^+,\cdots,\sigma_{n^-}^+)$ 是从 $\{-1,1\}$ 中均匀且独立同分布取的 Rademacher 随机变量，$\ell_{ij}=\ell(f(\boldsymbol{x}_i^+)-f(\boldsymbol{x}_j^-))$，$\mathcal{D}$是$\mathcal{S}$ 的分布。基于此，可得如定理 2 所示的抽象泛化界。然后，借助链式界技术，可推导出期望成对 Rademacher 复杂度的上界，从而得到定理 3 所示的泛化界。

定理 2. 抽象泛化界. 设 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ 为从分布 $\mathcal{D}$ 中独立同分布采样得到的训练集。对于任意得分函数 $f\in\mathcal{F}$ 和 $\delta\in(0,1)$，下式以至少 $1-\delta$ 的概率成立：

$$\begin{aligned}\mathcal{R}^{\ell_{0,1}}(f)&\leqslant\frac{q}{2}\mathcal{R}^{\ell}(f,w)\\&\leqslant\frac{q^2}{2}\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})+\frac{2q^2\rho}{\epsilon}\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{H})\\&\quad+\frac{q^2\rho B_{\tilde{w}}B_\ell}{2\epsilon}\sqrt{\frac{2+\rho+1/\rho}{2n}\log\frac{1}{\delta}}。\end{aligned}$$

注记 1. 上述定理的证明依赖于稳定替代损失(引理 1)，详见附录 B。

定理 3. 在定理 2 的假设下，对于任意得分函数 $f\in\mathcal{F}$ 和 $\delta\in(0,1)$，下式以至少 $1-\delta$ 的概率成立：

$$\mathcal{R}^{\ell_{0,1}}(f)\leqslant\frac{q^2}{2}\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})+\tilde{O}(\sqrt{A/n\cdot\log(2/\delta)}),$$

其中，$\tilde{O}$ 是在大 O 复杂度符号的基础上忽略对数复杂度项。

注记 2. 上式第一项为本文提出的稳定 AP 经验替代风险，可通过选择适当的模型和训练技术充分降低。第二项为泛化误差，随着训练集规模的增加，以 $\tilde{O}(\sqrt{A/n\cdot\log(2/\delta)})$ 的收敛率降为零。综上所述，定理 3 保证了在适当的模型和数据下，可以通过最小化 $\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})$ 显著优化测试 AP 损失。

6 双层随机优化算法

6.1 算法描述

本小节描述求解优化目标 $(OP1)$ 的随机算法。为简化符号，记 $e(\boldsymbol{\theta};u)=h(\rho\cdot\mathbb{E}_{\boldsymbol{x}_i^+,\boldsymbol{x}_j^-}[v(\boldsymbol{x}_i^+;u)\cdot$

$\ell(\delta_{ij}^{+-})]$),其中 $v(\boldsymbol{x}_i^+;u)=\tilde{w}(\ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)-u)/B_\ell)$,则优化目标可改写为下列期望形式:

$$(OP2)\ \min_{\boldsymbol{\theta}} e(\boldsymbol{\theta};\mu^+),\ \mu^+=\mathbb{E}_{\boldsymbol{x}^+}[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)]。$$

上述优化目标函数显然为双层复合函数,内层为求解正例期望得分 μ^+。现有复合优化问题求解方法[54-55]通常假设内层期望与外层变量无关,但在(*OP*2)中,内层期望取决于外层优化变量 $\boldsymbol{\theta}$,易导致梯度估计有偏。为避免该特性导致的优化算法无法收敛,本文提出估计正样本的平均得分 $\mu_k^+=\mathbb{E}_{\boldsymbol{x}^+}[f_{\boldsymbol{\theta}_k}(\boldsymbol{x}^+)]$。具体地,记 μ_k^+ 和 e 的随机估计为 $\mu_k^+(\zeta_k)$ 和 $e(\boldsymbol{\theta};\cdot,\zeta_k)$,其中 ζ_k 为随机变量,通常受样本随机采样主导。受方差缩减技术启发[56],μ_k^+ 的估计 $\hat{\mu}_k$ 更新规则如下所示:

$$\hat{\mu}_k=\underbrace{(1-\lambda)\hat{\mu}_{k-1}+\lambda\mu_k^+(\zeta_k)}_{\text{指数滑动平均}}+\underbrace{(1-\lambda)(\mu_k^+(\zeta_k)-\mu_{k-1}^+(\zeta_k))}_{\text{方差缩减项}}。\tag{6}$$

结合滑动平均策略和方差缩减项,$\hat{\mu}_k$ 以快速收敛到 μ_k^+,从而避免有偏梯度估计的影响。

随后,模型参数 $\boldsymbol{\theta}$ 通过基于梯度的方法进行更新,其中损失使用 $\hat{\mu}_k$ 进行估计:

$$e(\boldsymbol{\theta}_k;\hat{\mu}_k,\zeta_k)=h(\mathbb{E}_{\boldsymbol{x}_i^+,\boldsymbol{x}_j^-}[\mathrm{sg}[v(\boldsymbol{x}_i^+;\hat{\mu}_k)]\cdot\ell(\delta_{ij}^{+-})]),$$

其中,$\mathbb{E}_{\boldsymbol{x}_i^+,\boldsymbol{x}_j^-}$ 表示在小批量样本上的平均操作,$sg[\cdot]$ 表示梯度截断操作。这是随着模型收敛,有 $\lambda\approx 0$ 和 $\boldsymbol{\theta}_k\approx\boldsymbol{\theta}_{k-1}$,此时 $\hat{\mu}_k$ 不能反传关于 $\boldsymbol{\theta}_k$ 的梯度。在这种情况下,用来自 $v(\boldsymbol{x}_i^+;\hat{\mu}_k)$ 的梯度更新 $\boldsymbol{\theta}$ 会导致平凡解:降低正样本得分 $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)$,使得 $v(\boldsymbol{x}_i^+;\hat{\mu}_k)$ 较小。这种平凡解与最大化正样本得分的目标相矛盾。因此,这里截断梯度以避免这种平凡解。

算法流程总结如算法 1 所示。在每次迭代中,我们首先用式(6)更新 $\hat{\mu}_k$(第 4 行)。然后,对采样的小批量样本计算 $v(\boldsymbol{x}_i^+;\hat{\mu}_k)$ 和 $\ell(\delta_{ij}^{+-})$(第 5－6 行)。在计算梯度后(第 7 行),使用一阶优化方法更新模型参数(第 8 行)。

算法 1. (*OP*1) 的随机优化.

输入:训练集 $\mathcal{S}=\mathcal{S}^+\cup\mathcal{S}^-$,最大迭代次数 T,学习率 α,动量 β,$\lambda\in(0,1)$,用于控制算法性质的常数 $s\in[0,1]$。

输出:模型参数 $\boldsymbol{\theta}_{T+1}$。

过程:

1. 初始化模型参数 $\boldsymbol{\theta}_1$,$\hat{v}_0$,$\boldsymbol{\gamma}_1$,$\boldsymbol{\gamma}_1^s$;

2. for $k=1$ to T do

3. 在随机变量 ζ_k 的控制下,从 $\mathcal{S}$ 中抽取小批量样本;

4. 使用式(6)更新 $\hat{\mu}_k$;

5. 在小批量中对所有正样本计算 $v(\boldsymbol{x}_i^+;\hat{\mu}_k)$;

6. 对所有正样本－负样本对 $\boldsymbol{x}_i^+,\boldsymbol{x}_j^-$ 计算 $\ell(\delta_{ij}^{+-})$;

7. 计算随机梯度 $g_k=\nabla e(\boldsymbol{\theta}_k;\hat{\mu}_k,\zeta_k)$;

8. 更新 $\gamma_{k+1}=\boldsymbol{\theta}_k-\alpha g_k$,$\gamma_{k+1}^s=\boldsymbol{\theta}_k-s\alpha g_k$,$\boldsymbol{\theta}_{k+1}=\gamma_{k+1}+\beta(\gamma_{k+1}^s-\gamma_k^s)$.

9. end for

6.2　收敛性分析

接下来,对提出的算法进行收敛性分析。本节收敛性证明依赖以下假设,即要求对于 $1\leqslant k\leqslant T$,任意 $\boldsymbol{\theta},\boldsymbol{\theta}',u,u'$,及 $a,b\in[0,B_{\tilde{w}}]$,下列条件均满足:

假设 1. 平滑性和 Lipschitz 连续性

$$\|\nabla e(\boldsymbol{\theta};\cdot,\zeta_k)-\nabla e(\boldsymbol{\theta}';\cdot,\zeta_k)\|\leqslant L_\theta\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|,$$
$$\|\nabla e(\cdot;u,\zeta_k)-\nabla e(\cdot;u',\zeta_k)\|\leqslant L_u\|u-u'\|,$$
$$\|f_{\boldsymbol{\theta}}(\cdot)-f_{\boldsymbol{\theta}'}(\cdot)\|\leqslant\phi_f\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|,$$
$$\|\tilde{w}(a)-\tilde{w}(b)\|\leqslant\phi_{\tilde{w}}\|a-b\|。$$

假设 2. 方差及梯度有界

$$\|\nabla e(\boldsymbol{\theta}_k;u)\|\leqslant G,$$
$$\mathbb{E}_{\zeta_k}[\|\nabla e(\boldsymbol{\theta};u)-\nabla e(\boldsymbol{\theta};u,\zeta_k)\|^2]\leqslant\kappa_e^2,$$
$$\mathbb{E}_{\zeta_k}[\|\mu_k^+-\mu_k^+(\zeta_k)\|^2]\leqslant\kappa_\mu^2.$$

假设 3. 无偏随机估计

$$\mathbb{E}_{\zeta_k}[\nabla e(\boldsymbol{\theta};u,\zeta_k)]=\nabla e(\boldsymbol{\theta};u),$$
$$\mathbb{E}_{\zeta_k}[\mu_k^+(\zeta_k)]=\mu_k^+。$$

注记 3. 假设 1 包含得分函数和加权函数的平滑性,常见神经网络模型均满足该条件。此外,假设 1 包含的平滑性条件是非凸优化分析的常用基本条件[57-58],与模型激活函数、替代损失等有关。近期研究[59]表明,通过增加残差连接,神经网络具备良好的近似平滑性。

注记 4. 假设 2 可通过选择合适的替代损失实现。由主流神经网络实现的得分函数满足良好的梯度有界性[57,60]。假设 3 由样本的随机均匀采样可得。

注记 5. 上述假设仅在收敛性分析中使用,前文的泛化分析是基于假设空间中所有模型均满足,无需上述针对特定模型的假设。

基于上述假设,本文给出定理 4 定量描述所提出 AP 优化算法的收敛速率(更多解释见附录 C):

定理 4. 算法 1 收敛率. 记 $\boldsymbol{\theta}^*$ 是最优参数，$B'=36C_a^2C_\lambda^{-1}L_u^2\phi_f^2\cdot[(s(1-\beta)-1)^2C_\beta^2+(1+s\beta)^2]$，$\Delta_e=(e(\boldsymbol{\theta}_1)-e(\boldsymbol{\theta}^*))$。设算法 1 的更新规则共执行 T 步，取 $\alpha=C_a/\sqrt{T}$，$\beta=\min\{1-1/\sqrt{2},1/\sqrt{3L_u},C_\beta/\sqrt{T}\}$，$\lambda=C_\lambda/\sqrt{T}$，则有

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\nabla e(\boldsymbol{\theta}_k;\widetilde{w}^e)^2\right]\leqslant O(T^{-1})+B'(G^2+\kappa_e^2)\cdot T^{-\frac{1}{2}}+2(1-\beta)C_a^{-1}\Delta_e\cdot T^{-\frac{1}{2}}+9L_u^2C_\lambda\kappa_\mu^2T^{-\frac{1}{2}}。$$

注记 6. 上述定理表明，算法 1 以 $O(1/\epsilon^4)$ 的速率收敛到一个 ϵ-驻点，这与主流一阶非凸问题优化器相同[58]。对比现有 AP 优化算法收敛性，Chen 等人[4-5]提出的误差驱动法仅适用于线性模型和线性可分数据，无法适配真实数据和深度学习模型。Qi 等人[61]所提出方法适用于一般非凸模型，以 $O(1/\epsilon^5)$ 的速率收敛，略慢于本方法。

注记 7. 收敛性主要由两项决定：

$$\underbrace{2(1-\beta)C_a^{-1}\Delta_e+B'G^2}_{\text{初始条件}}+\underbrace{B'\kappa_e^2+9L_u^2C_\lambda\kappa_\mu^2}_{\text{随机方差项}}$$

其中，第一项由初始条件控制。此外，假设随机方差 κ_e，κ_μ 随着批量大小的增加而减小[58-59]，则在较大的批量下收敛速度更快。然而，当 $\kappa_e\ll G$ 时，这种改进遇到边际效用，难以进一步提升收敛率。该结论与后文所述实验结果一致(见图 3 (c))。

6.3 理论结论与算法适用条件

根据第 5 节所述泛化性分析条件，本算法要求训练样本和测试数据服从独立同分布，以确保定理 2 和定理 3 中的集中不等式大概率成立。此外，根据定理 3，随着训练集规模 n 的增加，本文提出的稳定 AP 经验替代风险以 $\widetilde{O}(\sqrt{1/n})$ 的速率接近原始期望风险，因此本算法适用于具有充足训练样本的情况，对少样本学习的泛化性可能欠佳。

根据第 6 节所述收敛性分析条件，得分函数(通常由神经网络实现)应满足 Lipschitz 连续性、平滑性，对常见模型成立，但对于涉及不可微算子等组件的非典型架构，上述理论可能不适用。

此外，本文聚焦于 AP 优化领域常见的二分类设定，要求优化目标可拆解为若干二分类子问题，例如图像检索、目标检测等任务，暂未覆盖回归问题、标注噪声、连续标签等复杂设定。后续实验涉及的目标检测问题满足可拆解性，但由于锚框分配是基于预设的交并比阈值，存在一定噪声，对应理论分析超出本文范畴，仅视为近似分析。

7 实 验

本节在图像检索、目标检测和长尾图像分类这三个视觉任务上验证了所提出的 AP 优化算法。

7.1 图像检索

7.1.1 实验设置

7.1.1.1 数据集

本文在三个不同领域、规模较大的标准图像检索基准数据集上评估所提出的 AP 优化方法。每个数据集遵循官方的测试集划分，并将其余数据按 9∶1 的比例划分为训练集和验证集。数据集的详细描述如下，统计信息见表 1。

(1)Stanford Online Products(SOP)[62]是商品检索数据集，包含从 eBay 上爬取的 120053 张图像。它由属于 12 个超类的 22634 个类别组成。图像按[63]中的方法划分测试集，并且进一步地从其余图像划分出每个超类的训练集和验证集，使得两者无重合部分。在测试阶段，对于每个查询图像，整个测试集所有图像被视为候选图库。

表 1 图像检索数据集的统计信息。

数据集	划分	图像数	类别数	不平衡比例	每类平均图像数
SOP	训练集	53555	10181	6.0	5.3
	验证集	5996	1137	6.0	5.3
	测试集	60502	11316	6.0	5.3
PKU VehicleID	训练集	101947	11847	71.0	8.6
	验证集	11399	1317	51.5	8.7
	测试集	40365	4800	59.0	8.2
iNaturalist	训练集	289930	5115	200.6	56.7
	验证集	35916	575	125.4	62.5
	测试集	136093	2452	125.4	55.5

(2)PKU VehicleID[64]是车辆检索数据集，包含 221736 张图像，共有 26267 个车辆类别。给定查询图像，本文按照标准测试方案使用三种候选图库[64]：①小型列表包含 800 个类别的 7332 张图像；②中型列表包含 1600 个类别的 12995 张图像；③大型列表包含 2400 个类别的 20038 张图像。

(3)iNaturalist[65]是野外自然物种识别的大规模不平衡图像数据集。按照现有研究[2]的做法，本文采用 2018 年版本的 iNaturalist 数据集进行图像检索，其包含 461,939 张图像，共有 8,142 个类别。这些类别属于植物、昆虫、鸟类、哺乳动物等 13 个超类。与 SOP 数据集类似，遵循标准的测试集划分，并将其余数据划分为训练集或验证集，测试集和验

证集中的类别未在训练集中出现过。测试方案与SOP数据集相同。

7.1.1.2　实现细节

(1)网络结构。特征提取器采用ResNet-50架构[66],并由ImageNet−1k[67]上预训练的模型权重初始化。其输入为RGB图像,输出为256维的嵌入向量。给定查询 $\boldsymbol{z}_q$ 和图库列表 $\{\boldsymbol{z}_i\}_{i=1}^n$ 的 L_2 归一化嵌入,其相似度得分由余弦相似度 $\boldsymbol{z}_q^\top \boldsymbol{z}_i$ 表示,其中 $i=1,2,\cdots,n$。

(2)深度特征混合(Deep Feature Mixup,DFM)。根据本文的泛化性分析,数据多样性和批量大小在AP优化的泛化中起着重要作用。鉴于此,引入类似混合[68]的数据增强方法来扩大批量大小。然而,对原始图像的数据增强和大批量都会带来较高的计算成本。本文认为,对原始图像进行数据增强不是必要的,因为低级信息的多样性不如高级语义信息重要。同时,主要的计算成本集中在浅层。

基于上述考虑,本文提出在深度特征上执行混合。具体来说,在每个小批量中,从同一个类中取出一组深度特征 $\mathcal{Z}=\{z_i\}_{i=1}^{|\mathcal{Z}|}$,并通过随机线性插值生成一组相同大小的增强样本 $\widetilde{\mathcal{Z}}=\{\tilde{\boldsymbol{z}}_i\}_{i=1}^{|\mathcal{Z}|}$:

$$\tilde{\boldsymbol{z}}_i = u\boldsymbol{z}_i + (1-u)\sum_{j=1}^{|\mathcal{Z}|} v_j \boldsymbol{z}_j,$$

其中,$v\sim U(0,1)$,u 是一个超参数,用于控制插值特征到原始特征的距离,在所有实验中设为0.9。之后,将 $\mathcal{Z}$ 和 $\widetilde{\mathcal{Z}}$ 合并后通过更深的层进行传递。显然,这种操作使得样本数量翻倍。本文在ResNet的最后两个阶段利用特征混合,这使得批量大小增加到原来的四倍。

(3)代理损失实现。考虑对异常样本的鲁棒性,本文采用由超参数 τ 控制的单边Huber损失:

$$\ell(x)=\begin{cases} -2x/\tau+1, & x<0,\\ (x/\tau-1)^2, & 0\leqslant x<\tau,\\ 0, & x\geqslant\tau. \end{cases} \tag{7}$$

代理函数具有以下属性:首先,ℓ 是 $\ell_{0,1}$ 的上界,且在 $f<\infty$ 时有 $\ell<\infty$,满足定理1的假设。此外,ℓ 是 $\frac{2}{\tau}$-平滑的、可微的并且导数处处连续,便于梯度优化方法的应用和分析;其次,与平方损失和指数损失相比,当 $x<0$ 时,$\ell(x)$ 是线性的,损失函数上界更为紧致。根据定理1分析结论,这使得泛化界更为紧致,泛化误差更小。

(4)优化策略。在训练阶段,图像尺寸调整为 256×256,通过标准数据增强方法扩充样本,包括随机翻转(50%的概率)、随机旋转(范围为[−15°,15°])和随机裁剪(224×224)。模型由算法1所示流程端到端更新,其中 $\lambda=0.01$,动量 $\beta=0.9$,并设置 $s=1$ 以启用Nesterov加速。加权函数中的超参数设置为 $a=0.1$,$t=2$。批量大小设置为112,单个小批量中每个类别恰好有4个正例。学习率 α 根据每个数据集在验证集上的表现进行调优。最优学习率如下:对于SOP数据集,学习率初始值为0.003,权重衰减为 4×10^{-5},在60 k和80 k次迭代时衰减0.1,最大迭代次数为100 k;对于VehicleID数据集,学习率初始值为0.01,权重衰减为 4×10^{-5},在120 k和150 k次迭代时衰减0.3,最大迭代次数为160 k;对于iNaturalist数据集,学习率初始值为0.001,权重衰减为 4×10^{-5},在150 k和220 k次迭代时衰减0.3,最大迭代次数为270 k。

(5)问题拆分。图像检索问题可视为给定查询样本作为条件的二分类问题。具体地,给定 Q 个查询 $\{q_k\}_{k=1}^Q$,候选图库根据是否与查询 q_k 相关划分为两个子集 $\{\boldsymbol{x}_i^k\}_{i=1}^{n_k}=\{\boldsymbol{x}_i^{k+}\}_{i=1}^{n_k^+}\cup\{\boldsymbol{x}_j^{k-}\}_{j=1}^{n_k^-}$,在模型训练阶段每次采样若干查询图像和候选图,优化目标为所采样查询的损失之和。

(6)评估指标。为验证AP优化性能,考虑使用mAP指标,即所有查询的AP的平均值。具体地,给定 Q 个查询 $\{q_k\}_{k=1}^Q$ 和相应的候选图库,特征提取器 m 的mAP指标如下:

$$\mathrm{mAP}(m)=\frac{1}{Q}\sum_{k=1}^{Q}\sum_{i=1}^{n_k^+}\frac{\sum_{j=1}^{n_k^+}\ell_{0,1}(f_k(\boldsymbol{x}_i^{k+})-f_k(\boldsymbol{x}_j^{k+}))}{\sum_{j=1}^{n_k}\ell_{0,1}(f_k(\boldsymbol{x}_i^{k+})-f_k(\boldsymbol{x}_j^{k}))},$$

其中,$f_k(x)=m(q_k)^\top m(x)$,满足 $m_2=1$。遵循以往的图像检索研究[2,12,15],本文还报告了Recall@K指标。注意这里的Recall@K定义与推荐系统中常用的不同。通常,图像检索中的Recall@K是指至少一个正例在前 K 名列表中排名的概率,具体形式为

$$Recall@K(m)=\frac{1}{Q}\sum_{k=1}^{Q}1\Big[\sum_{i=1}^{n_k^+}1[rank(\boldsymbol{x}_i^{k+})\leqslant K]>0\Big]。$$

7.1.1.3　对比方法

为验证所提出方法在图像检索中优势,本文选择三类算法作为对比基线:①基于成对样本损失的算法,包括对比损失(Contrastive loss)[69]、三元组损失(Triplet loss)[70]和多相似度(Multi-Similarity,MS)损失[71];②基于样本排序损失的算法,包括

SmoothAP[2]、Deep Image Retrieval (DIR)[15]、FastAP[3]、SoDeep[72]和 AUC [73]，其中除 AUC 外均针对 AP 指标优化；③黑盒指标优化方法[74]。

7.1.2 定量结果分析

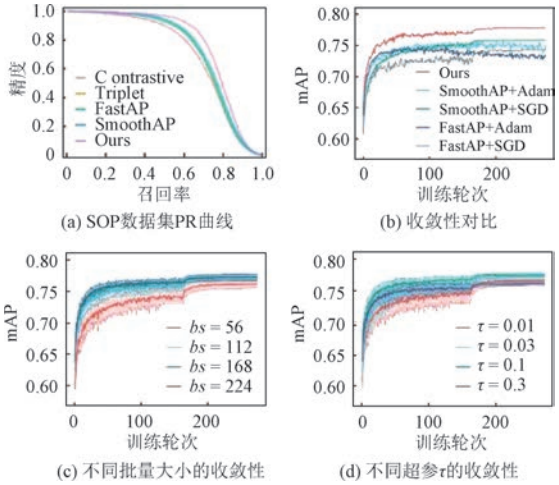
(1)主要结果：算法性能对比如表 2 所示，可从中观察得以下结论结果：①在所有的数据集上，本文所提出方法在 mAP 指标上明显优于所有对比方法。具体而言，所提出方法在 SOP、VehicleID 和 iNaturalist 数据集上分别比最佳基线方法高出 3.8%、0.9%和 1.2%。这验证了本文所提出稳定 AP 优化算法可以更有效地提升 AP 指标。此外，所提出算法的有效性在多个数据集上高度一致，验证了所提出的框架的应用潜力。②虽然在 SOP 和 VehicleID 数据集上，基于成对样本损失的方法在 Recall@1 上取得了令人满意的性能，但在其他指标上的结果相对较低。可以看出，成对样本损失更关注 Recall@1 或者与其等价的 Top-1 指标性能。相比之下，其他方法可以更好地平衡整体性能，特别是基于 AP 优化的方法。③基于样本排序的算法在大规模不平衡数据集 iNaturalist 上取得更为明显的性能提升。这表明 AP 优化更适合于严重不平衡的数据分布。

表 2 图像检索的定量结果

对比方法	SOP			PKU VehicleID						iNaturalist		
				小型目标		中型目标		大型目标				
	mAP	R@1	R@10	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	R@4
Contrastive loss[69]	55.50	75.59	88.09	69.05	79.92	61.28	74.81	54.42	68.44	23.41	49.05	67.04
Triplet loss [70]	57.94	78.26	90.16	79.51	93.56	75.38	92.20	71.86	90.81	27.67	56.27	73.91
MS loss[71]	57.67	77.33	98.59	70.14	80.86	62.74	76.33	56.25	70.59	25.94	53.70	71.02
SmoothAP[2]	**59.10**	**79.16**	**90.78**	79.90	94.12	76.46	92.72	72.08	91.12	**29.48**	**59.40**	**76.01**
DIR[15]	58.54	78.65	90.49	80.49	93.79	76.99	92.60	73.14	91.08	29.26	59.19	75.92
FastAP[3]	56.44	76.52	89.20	81.13	**94.39**	77.13	92.64	72.59	90.50	26.45	51.50	69.41
SoDeep[72]	55.51	76.40	88.94	61.00	74.57	53.26	68.47	46.38	61.50	20.42	45.32	64.72
AUC-Sigmoid[73]	55.96	76.91	89.68	75.28	90.93	70.70	89.00	65.39	85.65	24.29	56.20	74.30
AUC-Huber[73]	57.50	77.98	90.45	75.42	91.34	70.58	88.62	65.51	85.44	26.04	57.76	75.62
BlackBox[74]	58.19	78.14	89.95	**81.74**	94.04	**78.07**	**92.98**	**74.16**	**91.29**	27.64	54.45	72.14
Ours	**62.89**	**81.57**	**92.37**	**82.63**	**94.55**	**78.99**	**93.30**	**75.13**	**91.62**	**30.54**	**61.12**	**77.66**

注：所有模型均基于 ResNet-50 实现。最好的和次好的结果分别使用浅红色和浅蓝色标记。

(2)精度－召回率(Precision-Recall，PR)曲线：为了直观地从 PR 曲线的角度比较模型，图 3(a)中绘制了 SOP 数据集上的 PR 曲线，从中可以看出，本文提出的方法在精度和召回率之间做出了更好的权衡。



(a) SOP数据集PR曲线 (b) 收敛性对比

(c) 不同批量大小的收敛性 (d) 不同超参τ的收敛性

图 3 不同实验设置下的定性实验结果

(3)算法收敛率：图 3(b)中对比了 SmoothAP、FastAP 等方法的收敛速度，可以看出本文所提出方法具有更快的收敛速度。此外，图 3(c)和(d)展示了在不同超参数 τ（从 $\{0.01,0.03,0.1,0.3\}$ 中选择）和批量大小（从 $\{56,112,168,224\}$ 中选择）下的算法收敛速度，与本文的理论分析一致：定理 4 表明，增大批量可减小随机方差，但当随机方差 $\kappa_e \ll G$ 时，这种改进提升不再显著，与图中批量从 168 增大至 224 时提升不明显的现象一致；当 τ 充分小时，算法稳定性较差，对应模型泛化性能减弱，与定理 2 分析结果一致。

(4)算法复杂度分析：给定包含 n^+ 个正样本和 n^- 个负样本的批次，算法 1 所涉及主要计算包括得分计算(模型前向推断)、损失计算及模型反向传播，与最先进方法性能对比如表 3 所示。鉴于所有算法均采用同一架构，前向推断耗时相近。鉴于 DIR 损失复杂度为 $O(n^+ + n^-)$，而其他 AP 优化方法均为 $O(n^+ n^-)$，损失计算耗时相对 DIR 较高，但相较前向推理和反向传播增幅有限，对整体耗时影响仍可控。本方法复杂度与其他方法相同，由于涉及额

外步骤,耗时略有增加(不超过 10%),仍在可接受范围内。

表 3　算法耗时对比。

算法	模型前向推断	损失计算	模型反向传播	总计
SmoothAP[2]	246	40	528	814
DIR[15]	248	1.2	473	722
FastAP[3]	247	29	469	745
BlackBox[74]	248	40	472	760
Ours	248	57	534	839

7.1.3　消融实验

为了研究所提方法中不同技术的作用,本文在 SOP 和 iNaturalist 数据集上展开消融研究,结果如表 4 所示。

(1)正例得分估计 $\hat{v}_k$ 动态更新策略:如定理 4 所证明,以滑动平均方式估计平均正样本分数可以得到很好的收敛性。比较表 2 中的第 1 行和第 3 行、第 2 行和第 5 行可知,固定 $\hat{v}_k$ 会导致 mAP 在 SOP 和 iNaturalist 上分别降低 1.5%和 1.2%。这是由正例平均得分有偏估计所导致的,与本文的理论结果相符。

表 4　本文所提出算法中不同技术的消融实验结果

对照组编号	动态更新 $\hat{v}_k$	加入 DFM	t 取值	SOP					iNaturalist				
				mAP	R@1	R@10	R@100	R@1000	mAP	R@1	R@4	R@16	R@32
1			2	60.01	79.53	91.37	96.55	98.85	29.21	59.23	76.37	87.24	90.97
2		√	2	61.40	80.73	92.03	96.63	**98.88**	29.39	60.43	77.24	**87.86**	**91.51**
3	√		2	62.14	80.83	92.04	96.65	98.83	**30.68**	**60.81**	**77.24**	87.65	91.13
4	√	√	1	**62.73**	**81.35**	**92.34**	**96.73**	98.87	30.27	60.10	76.51	86.97	90.62
5	√	√	2	**62.89**	**81.57**	**92.37**	**96.81**	**98.90**	**30.54**	**61.12**	**77.66**	**87.93**	**91.46**

(2)深度特征混合(DFM):对比第 1 行和第 2 行、第 3 行和第 5 行可以看出,使用 DFM,除了在 iNaturalist 数据集上的 mAP 略有下降外,在多数对照组中有效提高整体性能,表明了 DFM 是一种有效的图像检索增强,并且值得进一步探索。

(3)加权函数设计:式(4)提供了受超参数 t 控制的一类加权函数原型,这里展示了 $t=1$ 和 $t=2$ 的结果。对比第 4 行和第 5 行,可以看出在 SOP 上的结果略有不同,而在 iNaturalist 上,将 t 从 1 改为 2 会带来明显的性能提升。这是因为较大的 t 值会导致更倾斜的权重分布,这与大规模数据集中的真实排序更一致。在实际应用中,t 值应根据数据规模进行调整。

7.2　目标检测

7.2.1　实验设置

7.2.1.1　数据集

下列所有实验均在大规模目标检测数据集 MS COCO[75] 数据集上进行,该数据集包含 165482 张图像,80 类目标,平均每张图像包含 7.7 个检测实例。采用默认设置,即所有模型都用 trainval135k 子集进行训练,并在 test-dev 数据集上进行测试。

7.2.1.2　实现细节

(1)网络架构。遵循以往的研究[4-5],本文使用 RetinaNet[76] 作为基础模型。网络骨干采用在 ImageNet 数据集上预训练的 ResNet-50 架构[66],其中包含一个 4 层的特征金字塔网络(Feature Pyramid Network,FPN),在每一层,锚定框只有一个尺度和一个比例 1:1。在训练阶段,批归一化层参数固定。

(2)损失函数。在典型的单阶段检测器中,有 C 个类别,给定锚定框集合 $A=\{a_i\}_{i=1}^{|A|}$,每个锚框将被分配一个标签 $y\in\{0,1,\cdots,C\}$,其中标签 0 表示背景。多分类问题以一对全部的方式分解,体来说,每个锚框 a_i 被复制 C 次以获得 $b_{i,j}$,其中 $j=1,2,\cdots,C$,相应的标签为 $y_{i,j}=1[y_i=j]$。这样,分类部分就变成了一个二分类问题,可以用 AP 优化方法解决。给定一个小批量的目标 $\{x_i\}^n$,检测器将输出几个边界框 $\{f^{reg}(x_i)\}_{i=1}^n$ 和分数 $\{f^{cls}(x_i)\}_{i=1}^n$。给定真实的边界框 $\{b_i\}$ 和标签 $\{y_i\}_{i=1}^n\in\{0,1\}^n$,总体损失如下:

$$\mathcal{L}_{ap}=h\left(\frac{1}{n}\sum_{y_i=1}\tilde{w}_i\ell_i\right),$$

$$\mathcal{L}_{reg}=\sum_{y_i=1}L_{\mathrm{GIoU}}(f^{reg}(x_i),b_i),$$

$$\mathcal{L}_{bias}=\left|\sum_{y_i=1}f^{cls}(\boldsymbol{x}_i)\right|^2,$$

$$\mathcal{L}_{det}=\mathcal{L}_{ap}+\lambda_1\mathcal{L}_{reg}+\lambda_2\mathcal{L}_{bias},$$

其中,$\mathcal{L}_{\mathrm{GIoU}}$ 指的是 GIoU 损失[77]。$\mathcal{L}_{bias}$ 用于对分数进行正则化,因为所提出的 AP 损失对分数偏差不

敏感，即给所有分数加上一个常数不会改变损失。这使得在后处理阶段确定阈值变得很麻烦，因此将正样本分数正则化为接近零。在所有实验中，取 $\lambda_1=2$，$\lambda_2=0.001$。

(3)优化策略。在训练阶段，图像尺寸调整为 512×512。类似于以往的研究[4,38,78]，采用 SSD[79] 中的标准数据增强方法。模型优化过程如算法 1 所示，其中目标函数被替换为 $\mathcal{L}_{det}$。超参数设置如下：$\lambda=0.01$，$\beta=0.9$，$s=0$，批量大小设置为 32(每块 GPU 8 张图像)。初始学习率设置为 0.008，在第 75 和 95 轮时衰减 0.1，总轮数设置为 100。此外，使用 ATSS[80] 作为锚定框分配器。

(4)问题拆分。目标检测任务涉及多个类别，在此根据类别拆分为若干二分类子问题。具体而言，给定其中一个类，本文根据预设阈值将候选框分为正样本和负样本，其中正样本指属于该类的候选框，负样本包括属于其他类的候选框和背景。最终优化目标为所有类对应二分类子问题的损失之和。

(5)评估指标。测试时，图像尺寸调整为短边不超过 500 像素。对得分最高的前一千个输出使用 IoU 阈值为 0.6 的非极大值抑制(Non-maximum suppression，NMS)。根据 MS COCO[48] 的官方评估指标，本文报告了 IoU 阈值为 0.5 和 0.75 时的 mAP(分别记作 AP_{50} 和 AP_{75})，以及阈值为 0.5 至 0.95 之间的 10 个阈值下的平均值(记作 mAP)。此外，还报告了小型、中型和大型目标的 AP，分别记作 AP_S、AP_M 和 AP_L。

7.2.1.3 对比方法

首先，本文比较了目标检测领域常用的 Focal Loss[76] 作为基线模型。为了验证所提出方法在目标检测中相对于其他基于排序方法的优势，还比较了以下基线方法：

(1)AP Loss[4-5]。Chen 等人提出以误差驱动的方式解决 AP 优化问题，即通过链式法则和有限差分近似，避免代理损失近似误差导致的 AP 优化误差。

(2)DR Loss[80]。Qian 等人通过分布式排序近似处理分类问题，分离正负样本的分布。

(3)aLRP Loss[38]。在 AP Loss 的基础上，Oksuz 等人将误差驱动技术引入回归损失中。

此外，本文还比较了最先进的两阶段检测器 Faster RCNN[81](使用交叉熵损失)以及两个流行的单阶段基线模型 SSD[79] 和 YOLOv3[81]。

7.2.2 定量结果分析

主要结果如表 5 所示。为了公平比较，表中报告了骨干网络和图像尺度基本一致下的模型性能。由表中可观察得以下结论：(1)对于单阶段检测器，基于排序的损失函数明显优于逐实例的损失函数(例如 Focal Loss)。这验证了基于排序的损失函数能更好地处理单阶段检测器中的不平衡问题。(2)在基于排序的损失函数中，本文所提出的稳定 AP 优化方法显示出更先进的性能，特别是在小目标检测方面。(3)使用所提出的方法，单阶段检测器可以达到与两阶段方法相媲美的性能。

表 5 MS COCO 测试集上的目标检测定量结果(%)

	方法	骨干网络	训练图像大小	测试图像大小	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
两阶段	FasterR-CNN+CE [81]	ResNet-50	500×500	500×500	39.10	59.30	42.40	**23.20**	43.40	53.50
	FasterR-CNN+aLRP [38]	ResNet-50	500×500	500×500	**41.00**	**61.10**	**43.70**	22.20	**44.90**	54.20
	FasterR-CNN+Ours	ResNet-50	500×500	500×500	41.20	62.10	44.40	23.80	45.40	**54.10**
单阶段	SSD [79]	VGG-16	513×513	513×513	28.80	48.50	30.30	10.90	31.80	43.50
	YOLOv3[82]	DarkNet-53	608×608	608×608	33.00	57.90	34.40	18.30	35.40	41.90
	RetinaNet+Focal [76]	ResNet-50	500×500	500×500	39.90	57.80	**43.30**	**22.20**	**44.10**	51.20
	APLoss [4-5]	ResNet-50	500×500	500×500	38.90	58.80	41.90	19.40	41.70	52.20
	DRLoss [80]	ResNet-50	500×500	500×500	37.40	56.00	40.00	20.80	41.20	50.50
	aLRPLoss [38]	ResNet-50	500×500	500×500	**40.50**	**59.60**	43.30	20.60	43.30	54.30
	Ours	ResNet-50	500×500	500×500	41.40	59.90	44.80	23.20	45.60	**53.40**

7.3 长尾分类

7.3.1 实验设置

7.3.1.1 数据集

实验所采用数据集包括两个常用长尾分类基准数据集 CIFAR-10-LT、CIFAR-100-LT[83]，分类包含 60000 张灰度图像，以及医学诊断基准数据集 ChestX-ray[84]，包含 172000 张 X 射线图像。遵循以往的长尾分类研究，本文将所有数据集划分为若干个二分类长尾数据集，并按 8∶1∶1 的比例划分为训练集、验证集和测试集。

7.3.1.2 实现细节

(1)网络架构。遵循现有研究的实验标准，对

CIFAR 系列数据集采用规模较小的 ResNet-18 模型作为骨干模型,对 ChestX-ray 数据集采用 ResNet-50 模型。模型输出经过 Sigmoid 函数映射至(0,1)后视为预测得分。

(2)优化策略。所有实验中均采用 SGD 作为优化器,超参数设置为 $\lambda=0.01$,$\beta=0.9$,$s=0$,学习率根据验证集上的结果调优。

(3)评估指标。在验证集上获得最优超参数后,更换随机种子重复所有实验 4 次,报告多次重复实验的 AP 指标平均值±标准差。

7.3.1.3　对比方法

本文比较了两种类型的对比方法。(1)基于交叉熵的逐实例损失函数,包括朴素交叉熵(Cross-Entropy,CE),类平衡交叉熵 (CB-CE)[85],Focal 损失[76]和 LDAM 损失[86]。(2)基于排序的损失函数,包括 MAUC[30]、SmoothAP[2]和 SOAP[61],它们直接优化基于排序的指标,其中 SmoothAP 和 SOAP 直接优化 AP 指标,是最主要的基准方法。

7.3.2　定量结果分析

如表 6 所示,基于排序的损失函数在整体上表现优于其他损失函数。其中,本文所提出的方法始终比最佳对比方法高出约 2%,展示了良好的跨任务泛化性,验证了稳定 AP 目标及其优化方法对泛化性能的影响。

表 6　长尾分类的定量结果

方法	CIFAR-10-LT	CIFAR-100-LT	ChestX-ray
CE	78.9±0.46	75.9±1.11	26.95±0.70
CB-CE[85]	82.6±0.48	81.9±0.65	28.36±0.05
Focal [76]	80.1±0.11	79.6±0.46	28.29±0.42
LDAM [86]	82.4±0.98	76.6±2.33	26.13±2.44
MAUC [30]	80.8±1.28	78.6±0.22	28.66±0.27
SmoothAP [2]	82.1±0.46	81.3±0.09	25.61±0.75
SOAP [61]	**84.7±0.34**	**83.1±1.15**	**29.80±0.42**
Ours	86.5±0.39	85.3±0.93	32.19±0.28

8　结　论

本文提出了一种用于平均精度(AP)优化的随机学习框架。为了确保泛化性能,旨在设计一个稳定的 AP 替代目标,受此启发,提出了一种加权成对损失形式的替代优化目标。为了优化该排序加权损失,本文采用了滑动平均策略更新内层参数,导出了一种一阶随机优化算法。面向所提出的替代优化目标和优化算法,本文研究了相应的理论性质。首先,为了验证所提出的替代目标的有效性,本文证明了这种替代目标具有良好的稳定性,进而大概率能够带来良好的泛化性能。其次,本文证明了对于非凸模型,所提出的算法在充分训练的情况下能够收敛到局部最优。在包含 3 个任务、7 个基准数据集上的实验验证了本文所提出框架在提升 AP 指标泛化性的优势及其应用潜力。

参　考　文　献

[1] Davis J, Goadrich M. The relationship between precision-recall and roc curves//Proceedings of the International Conference on Machine Learning. Pittsburgh, USA, 2006: 233-240

[2] Brown A, Xie W, Kalogeiton V, et al. Smooth-ap: Smoothing the path towards large-scale image retrieval//Proceedings of the European Conference on Computer Vision. Online, 2020: 677-694

[3] Cakir F, He K, Xia X, Kulis B, Sclaroff S. Deep metric learning to rank//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1861-1870

[4] Chen K, Li J, Lin W, et al. Towards accurate one-stage object detection with ap-loss//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5119-5127

[5] Chen K, Lin W, See J, et al. Ap-loss for accurate one-stage object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 3782-3798

[6] Guo H, Zheng K, Fan X, et al. Visual attention consistency under image transforms for multi-label image classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 729-739

[7] Wang J, Yang Y, Mao J, et al. Cnn-rnn: A unified framework for multi-label image classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2285-2294

[8] Goadrich M, Oliphant L, Shavlik J. Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. Machine Learning, 2006, 64(1-3): 231-261

[9] Metzler D, Croft W B. A markov random field model for term dependencies//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, 2005: 472-479

[10] Mohapatra P, Jawahar C V, Kumar M P. Efficient optimization for average precision svm//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2014: 2312-2320

[11] Henderson P, Ferrari V. End-to-end training of object class detectors for mean average precision//Proceedings of the Asian Conference on Computer Vision. Taipei, China, 2016: 198-213

[12] Li Z, Min W, Song J, et al. Rethinking the optimization of average precision: Only penalizing negative instances before positive ones is enough. arXiv preprint arXiv: 2102.04640, 2021

[13] Mohapatra P, Rolinek M, Jawahar C V, et al. Efficient optimization for rank-based loss function//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3693-3701

[14] Qin T, Liu T, Li H. A general approximation framework for direct optimization of information retrieval measures. Information Retrieval, 2010, 13(4):375-397

[15] Revaud J, Almazan J, Rezende R S, et al. Learning with average precision: Training image retrieval with a listwise loss//Proceedings of the International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5107-5116

[16] Ustinova E, Lempitsky V. Learning deep embeddings with histogram loss//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 4177-4185

[17] Burges C, Ragno R, Le Q. Learning to rank with nonsmooth cost functions//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2006: 193-200

[18] Burges C J C. From ranknet to lambdarank to lambdamart: An overview. Learning, 2010, 23(11): 81-100

[19] Cao Z, Qin T, Liu T, et al. Learning to rank: From pairwise approach to listwise approach//Proceedings of the International Conference on Machine Learning. Corvallis, USA, 2007: 129-136

[20] Qin T, Zhang X, Tsai M, et al. Query-level loss functions for information retrieval. Information Processing & Management, 2008, 44(2):838-855

[21] Xia F, Liu T, Wang J, et al. Listwise approach to learning to rank: theory and algorithm//Proceedings of the International Conference on Machine Learning. Helsinki, Finland, 2008: 1192-1199

[22] Manning C, Schutze H. Foundations of Statistical Natural Language Processing. USA: MIT Press, 1999

[23] Herschtal A, Raskutti B. Optimising area under the roc curve using gradient descent//Proceedings of the International Conference on Machine Learning. Banff, Canada, 2004: 385-392

[24] Joachims T. A support vector method for multivariate performance measures//Proceedings of the International Conference on Machine Learning. Bonn, Germany, 2005: 377-384

[25] Calders T, Jaroszewicz S. Efficient auc optimization for classification//Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery. Warsaw, Poland, 2007: 42-53

[26] Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003, 4: 933-969

[27] Lei Y, Ying Y. Stochastic proximal auc maximization. Journal of Machine Learning Research, 2021, 22(61): 1-45

[28] Liu M, Zhang X, Chen Z, et al. Fast stochastic auc maximization with $o(1/n)$-convergence rate//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 3189-3197

[29] Natole M, Ying Y, Lyu S. Stochastic proximal algorithms for auc maximization//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 3710-3719

[30] Yang Z, Xu Q, Bao S, et al. Learning with multiclass auc: Theory and algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11):7747-7763

[31] Zhao P, Hoi S C H, Jin R, et al. Online auc maximization//Proceedings of the International Conference on Machine Learning. Bellevue, USA, 2011: 233-240

[32] Ying Y, Wen L, Lyu S. Stochastic online auc maximization//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 451-459

[33] Raghavan V, Bollmann P, Jung G S. A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems, 1989, 7(3):205-229

[34] Boyd K, Eng K H, Page C D. Area under the precision-recall curve: point estimates and confidence intervals//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Prague, Czech, 2013 451-466

[35] Wei X, Shen Y, Sun X, et al. Attribute-aware deep hashing with self-consistency for large-scale fine-grained image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13904-13920

[36] Wei X, Xu S, Chen H, et al. Prototype-based classifier learning for long-tailed visual recognition. Science China Information Sciences, 2020, 65(6): 160105

[37] Yue Y, Finley T, Radlinski F, et al. A support vector method for optimizing average precision//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, The Netherlands, 2007: 271-278

[38] Oksuz K, Cam B C, Akbas E, et al. A ranking-based, balanced loss function unifying classification and localisation in object detection//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2020: 15534-15545

[39] Huang C, Zhai S, Guo P. Metricopt: Learning to optimize black-box evaluation metrics//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 174-183

[40] Jiang Q, Adigun O, Narasimhan H, et al. Optimizing black-box metrics with adaptive surrogates//Proceedings of the International Conference on Machine Learning. Vienna, Aus-

tria, 2020: 4784-4793

[41] Vlastelica M, Paulus A, Musil V, et al. Differentiation of blackbox combinatorial solvers. arXiv preprint arXiv: 1912.02175, 2019

[42] Rol′ınek M, Musil V, Paulus A, et al. Optimizing rank-based metrics with blackbox differentiation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7620-7630

[43] Bartlett P L, Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 2002, 3(Nov): 463-482

[44] Poggio T, Shelton C R. On the mathematical foundations of learning. American Mathematical Society, 2002, 39 (1): 1-49

[45] Valiant L G. A theory of the learnable. Communications of the ACM, 1984, 27(11): 1134-1142

[46] Usunier N, Amini M R, Gallinari P. Generalization error bounds for classifiers trained with interdependent data//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2005: 1369-1376

[47] Lan T, Liu T, Ma Z, et al. Gen eralization analysis of listwise learning-to-rank algorithms//Proceedings of the International Conference on Machine Learning. Montréal, Canada, 2009: 577-584

[48] Chen W, Liu T, Ma Z. Two-layer generalization analysis for ranking using rademacher average//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 370-378

[49] Tewari A, Chaudhuri S. Generalization error bounds for learning to rank: Does the length of document lists matter? //Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 315-323

[50] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology, 1975, 12(4):387-415

[51] Gao W, Jin R, Zhu S, et al. Onepass auc optimization//Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 906-914

[52] Gultekin S, Saha A, Ratnaparkhi A, et al. Mba: mini-batch auc optimization. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5561-5574

[53] Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. MIT Press, 2018

[54] Wang M, Liu J, Fang E. Accelerating stochastic composition optimization//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 1714-1722

[55] Wang M, Fang E X, Liu H. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. Mathematical Programming, 2017, 161(1-2):419-449

[56] Johnson R, Zhang T. Accelerating stochastic gradient descent using predictive variance reduction//Proceedings of the Advances in Neural Information Processing Systems. Stateline, USA, 2013

[57] Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization//Proceedings of the In International Conference on Machine Learning. Los Angeles, USA, 2019: 242-252

[58] Chen X, Liu S, Sun R, et al. On the convergence of a class of adam-type algorithms for non-convex optimization//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018

[59] Liu H, Chen M, Er S, et al. Benefits of overparameterized convolutional residual networks: Function approximation under smoothness constraint//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 13669-13703

[60] Zou D, Gu Q. An improved analysis of training over-parameterized deep neural networks//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2019: 2055-2064

[61] Qi Q, Luo Y, Xu Z, et al. Stochastic optimization of areas under precision-recall curves with provable convergence//Proceedings of the Advances in Neural Information Processing Systems. Online, 2021: 1752-1765

[62] Song H O, Xiang Y, Jegelka S, et al. Deep metric learning via lifted structured feature embedding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4004-4012

[63] Wang X, Hua Y, Kodirov E, et al. Ranked list loss for deep metric learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5207-5216

[64] Liu H, Tian Y, Wang Y, et al. Deep relative distance learning: Tell the difference between similar vehicles//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2167-2175

[65] Horn G V, Aodha O M, Song Y, et al. The inaturalist species classification and detection dataset//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8769-8778

[66] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[67] Russakovsky O, Deng J, Su H, et al. Imagenet large-scale visual recognition challenge. International Journal of Computer Vision, 115(3):211-252, 2015

[68] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization. arXiv preprint arXiv: 1710.09412, 2017

[69] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping//Proceedings of the IEEE/

CVF Conference on Computer Vision and Pattern Recognition. 2006:1735-1742

[70] Hoffer E, Ailon N. Deep metric learning using triplet network//Proceedings of the International Workshop on Similarity-Based Pattern Recognition. 2015:84-92

[71] Wang X, Han X, Huang W, et al. Multi-similarity loss with general pair weighting for deep metric learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5022-5030

[72] Engilberge M, Chevallier L, Perez P, et al. Sodeep: A sorting deep net to learn ranking loss surrogates//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10792-10801

[73] Gao W, Zhou Z. On the consistency of auc pairwise optimization//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 939-945

[74] Pogancic M V, Paulus A, Musil V, et al. Differentiation of blackbox combinatorial solvers//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-8

[75] Lin T, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[76] Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 2980-2988

[77] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 2019: 658-666

[78] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection//Proceedings of the International Conference on Computer Vision. Seoul, Korea (South), 2019: 6569-6578

[79] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 21-37

[80] Qian Q, Chen L, Li H, et al. Dr loss: Improving object detection by distributional ranking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2020: 12164-12172

[81] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2015: 91-99

[82] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018

[83] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Department of Computer Science, University of Toronto, Toronto, Canada, 2009: 1-60

[84] Wang X, Peng Y, Lu L, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2097-2106

[85] Cui Y, Jia M, Lin T, et al. Class-balanced loss based on effective number of samples//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9268-9277

[86] Cao K, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 1567-1578

[87] Ledoux M, Talagrand M. Probability in Banach Spaces: Isoperimetry and Processes. Berlin, Germany: Springer Science & Business Media, 2013

[88] Lei Y, Ding L, Bi Y. Local rademacher complexity bounds based on covering numbers. Neurocomputing, 2016, 218: 320-330

[89] Long P M, Sedghi H. Generalization bounds for deep convolutional neural networks//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-8

附录 A. 替代目标函数的性质

引理 2. 记 $f(x)=\dfrac{\sqrt{(1+x)x}}{a+x}$,$0<x\leqslant B$,$0<a\leqslant 1$,可以得到:

$$f(x)\leqslant\begin{cases}\dfrac{1}{2}(a-a^2)^{-1/2},0<a\leqslant\dfrac{1}{2},\\(1+B)/(a+B),\dfrac{1}{2}<a\leqslant 1。\end{cases}$$

证明. 求 $f(x)$ 对 x 的导数

$$f'(x)=\frac{(2a-1)x+a}{2(1+ax)\sqrt{(1+x)x}},$$

$$f'(x)=0\Rightarrow x=\frac{a}{1-2a}。$$

因此,当 $\dfrac{1}{2}<a\leqslant 1$ 时,$f(x)$ 单调递增,故可得

$$f(x)\leqslant f(B)=\frac{\sqrt{(1+B)B}}{a+B}\leqslant\frac{1+B}{a+B}。$$

当 $0<a\leqslant\dfrac{1}{2}$ 时,可得

$$f(x)\leqslant f\left(\frac{a}{1-2a}\right)=\frac{1}{2\sqrt{a-a^2}}。$$

证毕。

引理 3. 记 $r_i=\sum\limits_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^+)/n^+$,

$$w(r_i)=\begin{cases}\dfrac{1}{2}(r_i-r_i^2)^{-1/2},0<r_i\leqslant\dfrac{1}{2},\\(1+\rho B_\ell)/(r_i+\rho B_\ell),\dfrac{1}{2}<r_i\leqslant 1,\end{cases}$$

则可得

$$n^+\leqslant\sum_{i=1}^{n^+}w(r_i)\leqslant\left(\frac{\pi}{4}+\frac{1+\rho B_\ell}{1+2\rho B_\ell}\right)n^+。$$

证明. 不失一般性地,假设 n^+ 是偶数。为了证明不等式左侧部分,注意到 $w(r_i)$ 在 $\left(0,\dfrac{1}{2}\right]$ 和 $\left(\dfrac{1}{2},1\right]$ 上是单调递增的,故可得

$$\sum_{i=1}^{n^+}w(r_i)\geqslant\frac{n^+}{2}w\left(\frac{1}{2}\right)+\frac{n^+}{2}w(1)=n^+。$$

对于不等式右侧部分,注意到 $\left\{\sum\limits_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^+)/n^+\right\}_{i=1}^{n^+}$ 是 $\{1,2,\cdots,n^+\}$ 的一个排列,因此:

$$\sum_{i=1}^{n^+}w(r_i)=\sum_{i=1}^{n^+}w\left(\frac{1}{n^+}\right)=\frac{n^+}{2}\sum_{i=1}^{n^+/2}\frac{1}{n}\left[\left(1-\frac{i}{n^+}\right)\frac{i}{n^+}\right]^{-1/2}+\sum_{i=n^+/2+1}^{n^+}\frac{1+\rho B_\ell}{i/n^++\rho B_\ell}$$

$$\leqslant\frac{n^+}{2}\sum_{i=1}^{n^+/2}\frac{1}{n}\left[\left(1-\frac{i}{n^+}\right)\frac{i}{n^+}\right]^{-1/2}+\frac{n^+}{2}\cdot\frac{1+\rho B_\ell}{1/2+\rho B_\ell}。$$

注意到:

$$\frac{n^+}{2}\sum_{i=1}^{n^+/2}\frac{1}{n}\left[\left(1-\frac{i}{n^+}\right)\frac{i}{n^+}\right]^{-1/2}\leqslant\frac{n^+}{2}\int_0^{\frac{1}{2}}[(1-x)x]^{-1/2}\mathrm{d}x$$

$$=\frac{n^+}{2}\int_{\pi/2}^{\pi/4}[(1-\cos^2(t))\cos^2(t)]^{-1/2}(-2\cos(t)\sin(t))\mathrm{d}t=\frac{\pi n^+}{4}。$$

结合上述不等式,可以得到:

$$\sum_{i=1}^{n^+}w(r_i)\leqslant\left(\frac{\pi}{4}+\frac{1+\rho B_\ell}{1+2\rho B_\ell}\right)n^+。$$

证毕。

定理 1. 设 $\ell:\mathbb{R}\mapsto\mathbb{R}^+$ 是一个光滑函数,使得对于任意 $x\in\mathbb{R}$ 并且 $\ell\in[0,B_\ell]$,都有 $\ell_{0,1}(x)\leqslant\ell(x)$。记 $\rho=n^-/n^+$,$\tilde{h}(x)=\frac{x}{1+x}$,$h(x)=(\epsilon^2+\tilde{h}(x))^{-1/2}$,$\epsilon>0$,以及 $q=\frac{\pi}{2}+\frac{2(1+\rho B_\ell)}{1+2\rho B_\ell}$。可以得到:

$$\hat{\mathcal{R}}_{\mathcal{S}}^{\ell_{0,1}}(f)\leqslant\tilde{\mathcal{R}}_{\mathcal{S}}^{\ell}(\mathrm{f})\leqslant\frac{q}{2}\cdot h\left(\frac{\rho}{n^+}\sum_{i=1}^{n^+}w(r_i)\ell_i\right)\triangleq\frac{q}{2}\cdot\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f,w),$$

$$w(r_i)=\begin{cases}\frac{1}{2}(r_i-r_i^{\,2})^{-1/2},0<r_i\leqslant\frac{1}{2},\\(1+\rho B_\ell)/(r_i+\rho B_\ell),\frac{1}{2}<r_i\leqslant1。\end{cases}$$

证明. 首先对于式(1)中的 $\hat{\mathcal{R}}_{\mathcal{S}}^{\ell_{0,1}}(f)$:

$$\hat{\mathcal{R}}_{\mathcal{S}}^{\ell_{0,1}}(f)\leqslant\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}\left(\frac{\sum_{j=1}^{n^-}\ell(\delta_{ij}^{+-})}{\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^{+})}\right)=\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}(\rho\cdot\ell_i/r_i)\leqslant\frac{1}{n^+}\sum_{i=1}^{n^+}\frac{\sqrt{(1+\rho\ell_i)\rho\ell_i}}{r_i+\rho\ell_i}\cdot\sqrt{\frac{\rho\ell_i}{1+\rho\ell_i}+\epsilon^2}$$

$$\overset{\text{(a)}}{\leqslant}\frac{1}{n^+}\sum_{i=1}^{n^+}w(r_i)h(\rho\ell_i)\leqslant\frac{\sum_{i=1}^{n^+}w(r_i)}{n^+}\sum_{i=1}^{n^+}\frac{w(r_i)}{\sum_{i=1}^{n^+}w(r_i)}h(\rho\ell_i)$$

$$\overset{\text{(b)}}{\leqslant}\frac{\sum_{i=1}^{n^+}w(r_i)}{n^+}h\left(\rho\sum_{i=1}^{n^+}\frac{w(r_i)}{\sum_{i=1}^{n^+}w(r_i)}\ell_i\right)\overset{\text{(c)}}{\leqslant}\frac{q}{2}\cdot h\left(\rho\sum_{i=1}^{n^+}\frac{w(r_i)}{n^+}\ell_i\right)。$$

(a)请参考引理 2。

(b)由于 Jensen 不等式和 h 是凹函数。

(c)请参考引理 3。

证毕。

性质 1. 回顾 AP 经验风险:

$$\widehat{AP}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}(rank(\boldsymbol{x}_i^+,\mathcal{S}^+;f)/rank(\boldsymbol{x}_i^+,\mathcal{S};f))。$$

最大化上述风险相当于以下目标:

$$\min_{f\in\mathcal{F}}\hat{\mathcal{R}}_{\mathcal{S}}^{\ell_{0,1}}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}\left(\frac{\sum_{j=1}^{n^-}\ell_{0,1}(\delta_{ij}^{+-})}{\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^{+})}\right),$$

其中,$\tilde{h}(x)=x/(1+x)$ 是一个单调递增函数。

证明.

$$\widehat{AP}(f)=\frac{1}{n^+}\sum_{i=1}^{n^+}\frac{rank(\boldsymbol{x}_i^+,\mathcal{S}^+;f)}{rank(\boldsymbol{x}_i^+,\mathcal{S};f)}=\frac{1}{n^+}\sum_{i=1}^{n^+}\frac{\sum_{j=1}^{n^+}1[f(\boldsymbol{x}_i^+)<f(\boldsymbol{x}_j^+)]}{\sum_{j=1}^{n}1[f(\boldsymbol{x}_i^+)<f(\boldsymbol{x}_j)]}$$

$$=\frac{1}{n^+}\sum_{i=1}^{n^+}\frac{\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^{+})}{\sum_{j=1}^{n^-}\ell_{0,1}(\delta_{ij}^{+-})+\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^{+})}=1-\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{h}\left(\frac{\sum_{j=1}^{n^-}\ell_{0,1}(\delta_{ij}^{+-})}{\sum_{j=1}^{n^+}\ell_{0,1}(\delta_{ij}^{+})}\right)$$

证毕。

性质 2. 记

$$\widetilde{U}_{\mathcal{S}}^{\ell}(f)=\widetilde{h}\left(\frac{1}{n^{+}}\sum_{i=1}^{n^{+}}\rho/r_i\cdot\ell_i\right)\geqslant\widetilde{\mathcal{R}}_{\mathcal{S}}^{\ell}(f)。$$

对于任何满足 $0\leqslant\widetilde{w}\leqslant B$ 的加权函数和任意正常数 p ,存在一个 n^{+} 和一个得分函数,使得 $\widetilde{U}\ell_{\mathcal{S}}(f,w)>p\cdot\widetilde{U}_{\mathcal{S}}^{\ell}(f,\widetilde{w})$。

证明. 记 $w_i=w(r_i),\widetilde{w}_i=\widetilde{w}(r_i)$。根据 $\widetilde{h}$ 的单调性,我们只需证明 $\Delta=\sum_{i=1}^{n^{+}}(w_i\ell_i-p\widetilde{w}_i\ell_i)>0$。通过选择一个使得 $0<b_0\leqslant\ell_i\leqslant b_1$ 对所有 $1\leqslant i\leqslant n^{+}$ 都成立的评分函数,则有 $\Delta\geqslant\sum_{i=1}^{n^{+}}(w_ib_0-p\widetilde{w}_i\ell_i)\geqslant b_0\cdot n^{+}(1+\log(n^{+}))-Bpb_1\cdot n^{+}$。对于 $n^{+}>\exp(Bpb_1/b_0-1)$,可以得到 $\Delta>0$。

证毕。

附录 B. 泛化性分析

B. 1. 符号

为了便于表示,给定两个样本集合 $\{x_i^{+}\}_{i=1}^{n^{+}}\cup\{x_j^{-}\}_{j=1}^{n^{-}}$ 和 $\{x_i^{'+}\}_{i=1}^{n^{+}}\cup\{x_j^{'-}\}_{j=1}^{n^{-}}$,我们给出以下符号表示:

$$\begin{aligned}
\widetilde{w}_i&=\widetilde{w}(\ell(f(x_i^{+})-\mu^{+})/B_{\ell}),\\
\widetilde{w}_i^{'}&=\widetilde{w}(\ell(f(x_i^{'+})-\mu^{+})/B_{\ell}),\\
\ell_{ij}&=\ell(f(x_i^{+})-f(x_j^{-})),\\
\ell_{i'j}&=\ell(f(x_i^{'+})-f(x_j^{-})),\\
\ell_{ij'}&=\ell(f(x_i^{+})-f(x_j^{'-})),\\
\ell_{i'j'}&=\ell(f(x_i^{'+})-f(x_j^{'-})),\\
\ell_i&=\frac{1}{n^{-}}\sum_{j=1}^{n^{-}}\ell_{ij},\\
\ell'_i&=\frac{1}{n^{-}}\sum_{j=1}^{n^{-}}\ell_{i'j'}。
\end{aligned}$$

B. 2. 预备知识

引理 4. 有界差分不等式. 假设函数 f 满足具有常数 $c_1,c_2,\cdots,c_n$ 的有界差分假设,记为

$$v=\frac{1}{4}\sum_{i=1}^{n}c_i^{2}。$$

设 $Z=f(X_1,X_2,\cdots,X_n)$,其中 X_i 是独立的。则对于所有 $\lambda>0$,则可以得到

$$\log\mathbb{E}\left[\exp(\lambda(Z-\mathbb{E}(Z)))\right]\leqslant\frac{\lambda^{2}v}{2}。$$

引理 5. 极大值不等式. 设 $Z_1,Z_2,\cdots,Z_n$ 是满足方差因子为 v 的 sub-Gaussian 条件的实值随机变量,则可以得到

$$\mathbb{E}\left[\max_{i=1,\cdots,n}Z_i\right]\leqslant\sqrt{2v\log n}。$$

引理 6. Efron-Stein 不等式. 设 $X_1,X_2,\cdots,X_n$ 是独立随机变量,设 $Z=f(X_1,X_2,\cdots,X_n)$。如果 f 具有常数 $c_1,c_2,\cdots,c_n$ 的有界差分性质,则可以得到

$$Var(Z)\leqslant\frac{1}{4}\sum_{k=1}^{n}c_k^{2}。$$

引理 7. McDiarmid 不等式. 设 $X_1,X_2,\cdots,X_n$ 是独立随机变量,并且 $f:\mathcal{X}^{n}\mapsto\mathbb{R}$ 具有常数 $c_1,c_2,\cdots,c_n$ 的有界差分性质,那么对于任意 $\delta\in(0,1)$,至少以 $1-\delta$ 的概率有

$$f(X_1,X_2,\cdots,X_n)-\mathbb{E}[f(X_1,X_2,\cdots,X_n)]\leqslant\sqrt{\frac{1}{2}\log\frac{1}{\delta}\sum_{i=1}^{n}c_i^2}。$$

定义 3. ϵ-覆盖[87]. 设$(\mathcal{M},d)$是一个(伪)度量空间,$\mathcal{K}$是$\mathcal{M}$的一个子集。对于一个正数ϵ以及$\mathcal{H}=\{h_1,h_2,\cdots,h_n\}\subseteq\mathcal{M}$,如果$\mathcal{K}\subseteq\bigcup_{i=1}^{n}\mathcal{B}(h_i,\epsilon)$,则称$\mathcal{H}$为$\mathcal{K}$的一个$\epsilon$-覆盖,其中$\mathcal{B}(h,\epsilon)$表示以$h$为中心、以$\epsilon$为半径为的球。

定义 4. 覆盖数[87]. 设$(\mathcal{M},d)$是一个(伪)度量空间,$\mathcal{K}$是$\mathcal{M}$的一个子集。$\mathcal{K}$的半径为ϵ的覆盖数定义为:

$$\mathcal{N}(\mathcal{K},\epsilon,d)=\min\{n:\text{大小为 } n \text{ 的}\mathcal{K}\text{上存在}\epsilon\text{-覆盖}\}。$$

B. 3. 证明思路

首先,引入三个引理来帮助推导泛化界。

引理 8. 设f是从假设空间$\mathcal{F}$中选择的得分函数。假设损失函数ℓ的取值范围在$[0,B_\ell]$,则可以得到

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,w)-q\,\mathcal{R}^{\ell}(f,\tilde{w})]\leqslant 0,$$

其中,$q=\frac{\pi}{2}+\frac{2(1+\rho B_\ell)}{1+2\rho B_\ell}$。

引理 9. 设f是从假设空间$\mathcal{F}$中选择的得分函数。假设损失函数ℓ的取值范围在$[0,B_\ell]$,至少以$1-\delta$的概率有

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,\tilde{w})-\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f,\tilde{w})]\leqslant\frac{4\rho}{\epsilon}\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{G})+\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon}\sqrt{\frac{2+\rho+1/\rho}{2n}\log\frac{1}{\delta}}。$$

引理 10. 假设正负样本数量的比率$\rho=n^+/n^-$是固定的,$\tilde{w}\in[1,B_{\tilde{w}}]$,$\ell\in[0,B_\ell]$。至少以$1-\delta$的概率有

$$\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{G})\leqslant\mathfrak{R}^{p}_{\mathcal{S}}(\mathcal{G})+B_{\tilde{w}}B_\ell\sqrt{\frac{2(2+\rho+1/\rho)}{n}\log\frac{1}{\delta}},$$

其中,

$$\mathcal{G}=\{g(x_1,x_2;f,\tilde{w},\ell)=\tilde{w}(\ell(f(\boldsymbol{x}_1)-\mu^+)/B_\ell)\cdot\ell(f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2))\mid f\in\mathcal{F}\}。$$

定理 2. 抽象泛化界. 设$\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$为从分布$\mathcal{D}$中独立同分布采样得到的训练集。对于任意得分函数$f\in\mathcal{F}$和$\delta\in(0,1)$,以至少$1-\delta$的概率,可以得到

$$\mathcal{R}^{\ell_{0,1}}(f)\leqslant\frac{q}{2}\mathcal{R}^{\ell}(f,w)\leqslant\frac{q^2}{2}\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f,\tilde{w})+\frac{2q^2\rho}{\epsilon}\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{H})+\frac{q^2\rho B_{\tilde{w}}B_\ell}{2\epsilon}\sqrt{\frac{2+\rho+1/\rho}{2n}\log\frac{1}{\delta}}.$$

证明. 注意到

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,w)-\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f,\tilde{w})]\leqslant\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,w)-q\,\mathcal{R}^{\ell}(f,\tilde{w})]+q\cdot\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,\tilde{w})-\hat{\mathcal{R}}_{\mathcal{S}}^{\ell}(f,\tilde{w})],$$

根据引理 8 和引理 9,我们可以得到结论。

证毕。

定理 5. 链式界. 假设损失函数ℓ是ϕ_ℓ-Lipschitz 连续的,加权函数$\tilde{w}$是$\phi_{\tilde{w}}$-Lipschitz 连续的。假设得分函数f是B_ℓ-一致有界的,则可以得到以下结论:

(1)对于任意正常数$\eta\leqslant B_\ell$,可以得到

$$\mathfrak{R}^{p}_{\mathcal{S}}(\mathcal{G})\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta+\frac{12C_0}{\sqrt{n^+}}\int_{\eta}^{B_\ell}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d_\infty)}\,\mathrm{d}\epsilon,$$

其中,$C_0=\frac{2B_{\tilde{w}}\phi_\ell+B_\ell B_{\tilde{w}}}{2}\sqrt{2+\rho+1/\rho}$,

$d_\infty(f_1,f_2)=max_{x\in\mathcal{S}}|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|$。

(2)设假设类$\mathcal{F}$的覆盖数具有明确的形式:

$$\log(\mathcal{N}(\epsilon,\mathcal{F},d_\infty))\leqslant A\log\left(\frac{b}{\epsilon}\right),$$

取 $\eta=\frac{1}{\sqrt{n}}$,可以得到:

$$\mathfrak{R}_S^p(\mathcal{G})\leqslant\frac{2B_{\widetilde{w}}\phi_\ell+B_\ell\phi_{\widetilde{w}}}{\sqrt{n}}+\frac{12C_0B_\ell}{\sqrt{n}}\sqrt{\frac{A}{2}\log(B_\ell^2n)}。$$

注记 5. 在结论(2)中,覆盖数以一般形式表示,该形式适用于大量的假设类,如深度神经网络[88-89]。

结合定理 2、引理 9 和定理 5,定理 3 的证明完成。

B.4. 加权函数近似误差分析

引理 8. 设 f 是从假设空间 $\mathcal{F}$ 中选择的得分函数。假设损失函数 ℓ 的取值范围在 $[0,B_\ell]$,则可以得到:

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^\ell(f,w)-q\,\mathcal{R}^\ell(f,\widetilde{w})]\leqslant 0,$$

其中,$q=\frac{\pi}{2}+\frac{2(1+\rho B_\ell)}{1+2\rho B_\ell}$。

证明. 首先,证明对于任意 $k=1,2,\cdots,n^+$,都有 $\sum_{i=k}^{n^+}(w_i-q\widetilde{w}_i)\leqslant 0$。当 $k>n^+/2$ 时,可以得到:

$$\sum_{i=k}^{n^+}(w_i-q\widetilde{w}_i)=\sum_{i=k}^{n^+}\frac{(1+\rho B_\ell)\cdot n^+/i}{1+\rho B_\ell\cdot n^+/i}-q\sum_{i=k}^{n^+}\left(\frac{a+\rho B_\ell}{a+\rho\ell(x)}\right)^t$$
$$\leqslant\frac{2(1+\rho B_\ell)}{1+2\rho B_\ell}(n^+-k+1)-q(n^+-k+1)=-\frac{\pi}{2}(n^+-k+1)\leqslant 0。$$

当 $k\leqslant n^+/2$ 时,可以得到:

$$\sum_{i=k}^{n^+}(w_i-q\widetilde{w}_i)=\sum_{i=k}^{n^+}w_i-q\sum_{i=k}^{n^+}\left(\frac{a+B_\ell}{a+\ell(x)}\right)^t,$$

由引理 3

$$\leqslant\frac{q}{2}\cdot n^+-q(n^+-k+1)=q(k-1-n^+/2)<0。$$

记 $\ell_i=\frac{1}{n^-}\sum_{j=1}^{n^-}\ell(\delta_{ij}^{+-})$ 并且 $\{\ell_{(i)}\}_{i=1}^{n^+}$ 是 $\{\ell_i\}_{i=1}^{n^+}$ 的排列,使得当 $i<j$ 时 $\ell_{(i)}<\ell_{(j)}$。注意到 $0\leqslant\ell_i\leqslant B_\ell$ 以及当 $\ell_i>\ell_j$ 时有 $w_i<w_j$,因此,

$$\sum_{i=1}^{n^+}w_i\ell_i-q\sum_{i=1}^{n^+}\widetilde{w}_i\ell_i=\sum_{i=1}^{n^+}d_i\ell_{(i)},$$

其中,$d_i=w(n^+/i)-qw_i$。此外,可以得到:

$$\sum_{i=1}^{n^+}d_i\ell_{(i)}=\sum_{k=1}^{n}\sum_{i=k}^{n}d_i(\ell_{(k)}-\ell_{(k-1)})=\sum_{k=1}^{n}\left[(\ell_{(k)}-\ell_{(k-1)})\sum_{i=k}^{n}d_i\right]\leqslant 0,$$

其中,最后一个不等式因为 $(\ell_{(k)}-\ell_{(k-1)})\geqslant 0$ 以及 $\sum_{i=k}^{n}d_i\leqslant 0$。因此可以得到:

$$h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}w_i\ell_i\right)-qh\left(\frac{1}{n^+}\sum_{i=1}^{n^+}\widetilde{w}_i\ell_i\right)\leqslant h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}w_i\ell_i\right)-h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}q\widetilde{w}_i\ell_i\right)\leqslant 0。$$

由此可以推出:

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^\ell(f,w)-q\,\mathcal{R}^\ell(f,\widetilde{w})]=\sup_{f\in\mathcal{F}}\mathbb{E}_S[\hat{\mathcal{R}}_S^\ell(f,w)-q\,\hat{\mathcal{R}}_S^\ell(f,\widetilde{w})]$$
$$\leqslant\mathbb{E}_S\sup_{f\in\mathcal{F}}[\hat{\mathcal{R}}_S^\ell(f,w)-q\,\hat{\mathcal{R}}_S^\ell(f,\widetilde{w})]\leqslant 0。$$

证毕。

B.5. 经验风险和期望风险的差异

引理 9. 设 f 是从假设空间 $\mathcal{F}$ 中选择的得分函数。假设损失函数 ℓ 的取值范围在 $[0,B_\ell]$，至少以 $1-\delta$ 的概率有

$$\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,\tilde{w})-\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})]\leqslant\frac{4\rho}{\epsilon}\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{G})+\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon}\sqrt{\frac{2+\rho+1/\rho}{2n}\log\frac{1}{\delta}}。$$

证明. 记 $\Phi(\mathcal{S})=\sup_{f\in\mathcal{F}}[\mathcal{R}^{\ell}(f,\tilde{w})-\hat{\mathcal{R}}^{\ell}_{\mathcal{S}}(f,\tilde{w})]$，$\mathcal{S}_k=\mathcal{S}\backslash\{x_k\}\cup\{\boldsymbol{x}_k'\}$，我们证明 $\Phi(\cdot)$ 具备有界差异性质。注意到 h 是 $\frac{1}{2\epsilon}$-Lipschitz 连续的，因此可以得到以下条件：

(1)如果 $\boldsymbol{x}_k$ 是正样本，可以得到：

$$|\Phi(\mathcal{S})-\Phi(\mathcal{S}_k)|\leqslant\frac{\rho}{2\epsilon}\cdot\sup_{f\in\mathcal{F}}\left|\frac{1}{n^+n^-}\sum_{j=1}^{n^-}(\tilde{w}'_k\ell_{k'j}-\tilde{w}_k\ell_{kj})\right|\leqslant\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon n^+}。$$

(2)如果 $\boldsymbol{x}_k$ 是负样本，可以得到：

$$|\Phi(\mathcal{S})-\Phi(\mathcal{S}_k)|\leqslant\frac{\rho}{2\epsilon}\cdot\sup_{f\in\mathcal{F}}\left|\frac{1}{n^+n^-}\sum_{i=1}^{n^+}\tilde{w}_i\cdot(\ell_{ik'}-\ell_{ik})\right|\leqslant\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon n^-}。$$

可以由引理 7 直接得到，对于任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率，都有

$$\begin{aligned}\Phi(\mathcal{S})&\leqslant\mathbb{E}_{\mathcal{S}}[\Phi(\mathcal{S})]+\sqrt{\left[n^+\left(\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon n^+}\right)^2+n^-\left(\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon n^-}\right)^2\right]\cdot\frac{1}{2}\log\frac{1}{\delta}}\\&\leqslant\mathbb{E}_{\mathcal{S}}[\Phi(\mathcal{S})]+\frac{\rho B_\ell B_{\tilde{w}}}{\epsilon}\sqrt{\frac{2+\rho+1/\rho}{2n}\log\frac{1}{\delta}}\end{aligned}\tag{附 1}$$

接下来我们关注 $\mathbb{E}_{\mathcal{S}}[\Phi(\mathcal{S})]$。根据 $\mathcal{S}$ 和 $\mathcal{S}'$ 的对称性，可以得到：

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[\Phi(\mathcal{S})]&=\mathbb{E}_{\mathcal{S}}\sup_{f\in\mathcal{F}}\left[\mathbb{E}_{\mathcal{S}'}h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{w}_i'\ell_i'\right)-h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{w}_i\ell_i\right)\right]\\&\leqslant\mathbb{E}_{\mathcal{S},\mathcal{S}'}\sup_{f\in\mathcal{F}}\left[h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{w}_i'\ell_i'\right)-h\left(\frac{1}{n^+}\sum_{i=1}^{n^+}\tilde{w}_i\ell_i\right)\right]\\&\leqslant\frac{\rho}{2\epsilon}\cdot\mathbb{E}_{\mathcal{S},\mathcal{S}'}\sup_{f\in\mathcal{F}}\left|\frac{1}{n^+}\sum_{i=1}^{n^+}(\tilde{w}'_i\ell_i'-\tilde{w}_i\ell_i)\right|\\&=\frac{\rho}{2\epsilon}\cdot\mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n^+n^-}\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\left[+\left(\frac{\sigma_i+\sigma_j}{2}\right)(\tilde{w}'_i\ell_{i'j'}-\tilde{w}_i\ell_{ij})\right.\right.\\&\quad+\left(\frac{\sigma_i-\sigma_j}{2}\right)(\tilde{w}'_i\ell_{i'j}-\tilde{w}_i\ell_{ij'})-\left(\frac{\sigma_i+\sigma_j}{2}\right)(\tilde{w}'_i\ell_{ij}-\tilde{w}_i\ell_{i'j'})\\&\quad\left.\left.-\left(\frac{\sigma_i-\sigma_j}{2}\right)(\tilde{w}'_i\ell_{ij'}-\tilde{w}_i\ell_{i'j})\right]\right\}\leqslant\frac{2\rho}{\epsilon}\mathfrak{R}^{p}_{n^+,n^-}(\mathcal{G})\end{aligned}\tag{附 2}$$

其中，第二个不等式的证明类似于文献[30]中的引理 8。通过结合式(附 1)和式(附 2)完成本引理的证明。

证毕。

B.6. 链式界的证明

定理 5. 链式界. 假设损失函数 ℓ 是 ϕ_ℓ-Lipschitz 连续的，加权函数 $\tilde{w}$ 是 $\phi_{\tilde{w}}$-Lipschitz 连续的。假设得分函数 f 是 B_ℓ-一致有界的，则可以得到以下结论：

(1)对于任意正常数 $\eta\leqslant B_\ell$，可以得到：

$$\mathfrak{R}^{p}_{\mathcal{S}}(\mathcal{G})\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta+\frac{12C_0}{\sqrt{n^+}}\int_{\eta}^{B_\ell}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d_\infty)}\,\mathrm{d}\epsilon,$$

其中,$C_0=\dfrac{2B_{\widetilde{w}}\phi_\ell+B_\ell B_{\widetilde{w}}}{2}\sqrt{2+\rho+1/\rho}$,

$d_\infty(f_1,f_2)=\max\limits_{x\in\mathcal{S}}|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|$。

(2)设假设类 $\mathcal{F}$ 的覆盖数符合以下具体形式:

$$\log(\mathcal{N}(\epsilon,\mathcal{F},d_\infty))\leqslant A\log\left(\frac{b}{\epsilon}\right),$$

取 $\eta=\dfrac{1}{\sqrt{n}}$,可以得到:

$$\mathfrak{R}_{\mathcal{S}}^{p}(\mathcal{G})\leqslant\frac{2B_{\widetilde{w}}\phi_\ell+B_\ell\phi_{\widetilde{w}}}{\sqrt{n}}+\frac{12C_0B_\ell}{\sqrt{n}}\sqrt{\frac{A}{2}\log(B_\ell^2 n)}。$$

证明.首先,我们考虑 $\mathcal{F}$ 的一个 ϵ-覆盖,记为 $\hat{\mathcal{F}}$,它的一个(伪)度量为 d_∞。选择 $\hat{f}$, $\widehat{\widetilde{f}}\in\hat{\mathcal{F}}$,使得 $d_\infty(\hat{f},f)\leqslant\epsilon,d_\infty(\widehat{\widetilde{f}},\widetilde{f})\leqslant\epsilon$,然后可以得到:

$$\begin{aligned}\mathfrak{R}_{\mathcal{S}}^{p}(\mathcal{G})&=\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f,\widetilde{f}\in\mathcal{F}}|T_f(\boldsymbol{\sigma})-T_{\widetilde{f}}(\boldsymbol{\sigma})|\right]\\&=\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f,\widetilde{f}\in\mathcal{F}}|(T_f(\boldsymbol{\sigma})-T_{\hat{f}}(\boldsymbol{\sigma}))+(T_{\widehat{\widetilde{f}}}(\boldsymbol{\sigma})-T_{\widetilde{f}}(\boldsymbol{\sigma}))+T_{\hat{f}}(\boldsymbol{\sigma})-T_{\widehat{\widetilde{f}}}(\boldsymbol{\sigma})|\right]\\&\leqslant\underbrace{2\ \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{d(f,\hat{f})\leqslant\epsilon}|T_f(\boldsymbol{\sigma})-T_{\hat{f}}(\boldsymbol{\sigma})|\right]}_{(1)}+\underbrace{2\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\hat{f}\in\hat{\mathcal{F}}_\epsilon}|T_{\hat{f}}(\boldsymbol{\sigma})|\right]}_{(2)}\end{aligned}\tag{附 3}$$

考虑一个精确率递减的序列 $\{\eta_k\}_{k=1}^{\infty}$,其中 $\eta_0\geqslant B_\ell$,并且对于任意 $k=1,2,\cdots,K$ 都有 $\eta_{k+1}\geqslant\dfrac{1}{2}\eta_k$。设 $\hat{\mathcal{F}}_k$ 是 $\mathcal{F}$ 的一个 ϵ-覆盖,选择 $\hat{f}_k$, $\widehat{\widetilde{f}}_k$ 使得 $d_\infty(\hat{f}_k,f)\leqslant\eta_k,d_\infty(\widehat{\widetilde{f}}_k,\widetilde{f})\leqslant\eta_k$。特殊地,令 $\eta_K=\eta,\hat{f}_K=\hat{f}$, $\widehat{\widetilde{f}}_K=\widehat{\widetilde{f}}$。注意到 $T_{\hat{f}}(\boldsymbol{\sigma})=T_{\hat{f}_K}(\boldsymbol{\sigma})=T_{\hat{f}_0}(\boldsymbol{\sigma})+\sum\limits_{k=1}^{K}(T_{\hat{f}_k}(\boldsymbol{\sigma})-T_{\hat{f}_{k-1}}(\boldsymbol{\sigma}))$,则第(1)项可以被分解为

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{d_\infty(f,\hat{f})\leqslant\epsilon}|T_f(\boldsymbol{\sigma})-T_{\hat{f}}(\boldsymbol{\sigma})|\right]\\&\leqslant\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\hat{f}_k\in\hat{\mathcal{F}}_k,\hat{f}_{k-1}\in\hat{\mathcal{F}}_{k-1}\\d_\infty(\hat{f}_k,\hat{f}_{k-1})\leqslant3\eta_k}}|T_{\hat{f}_k}(\boldsymbol{\sigma})-T_{\hat{f}_{k-1}}(\boldsymbol{\sigma})|\right]\end{aligned}\tag{附 4}$$

根据引理 5 和引理 11,可以得到:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\hat{f}_k\in\hat{\mathcal{F}}_k,\hat{f}_{k-1}\in\hat{\mathcal{F}}_{k-1}\\d_\infty(\hat{f}_k,\hat{f}_{k-1})\leqslant3\eta_k}}|T_{\hat{f}_k}(\boldsymbol{\sigma})-T_{\hat{f}_{k-1}}(\boldsymbol{\sigma})|\right]&\leqslant C\sqrt{2\log|\hat{\mathcal{F}}_k||\hat{\mathcal{F}}_{k-1}|}\,d_\infty(\hat{f}_k,\hat{f}_{k-1})\\&\leqslant6\eta_kC\sqrt{\log(\mathcal{N}(\eta_k,\mathcal{F},d_\infty))}\end{aligned}\tag{附 5}$$

另一方面,第(2)项可以被上界约束:

$$\begin{aligned}&\mathbb{E}_\sigma\left[\sup_{d_\infty(f,\hat{f})\leqslant\eta_K}|T_f(\boldsymbol{\sigma})-T_{\hat{f}}(\boldsymbol{\sigma})|\right]=\\&\mathbb{E}_\sigma\left[\sup_{d_\infty(f,\hat{f})\leqslant\eta_K}\left|\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\left[\frac{\sigma_i^+-\sigma_j^-}{2n^+n^-}(\widetilde{w}_i^f\ell_{ij}^f-\widetilde{w}_i^{\hat{f}}\ell_{ij}^{\hat{f}})\right]\right|\right]\\&\leqslant\sup_{d_\infty(f,\hat{f})\leqslant\eta_K}\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}\frac{\widetilde{w}_i^f\ell_{ij}^f-\widetilde{w}_i^{\hat{f}}\ell_{ij}^f+\widetilde{w}_i^{\hat{f}}\ell_{ij}^f-\widetilde{w}_i^{\hat{f}}\ell_{ij}^{\hat{f}}}{n^+n^-}\end{aligned}$$

$$\leqslant \frac{1}{n^+n^-}\sup_{d_\infty(f,\hat{f})\leq\eta_K}\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}(|\tilde{w}_i^f\ell_{ij}^f-\tilde{w}_i^{\hat{f}}\ell_{ij}^{\hat{f}}|+|\tilde{w}_i^{\hat{f}}\ell_{ij}^f-\tilde{w}_i^{\hat{f}}\ell_{ij}^{\hat{f}}|)$$

$$\leqslant \frac{B_\ell}{n^+}\sup_{d_\infty(f,\hat{f})\leq\eta_K}\sum_{i=1}^{n^+}|\tilde{w}_i^f-\tilde{w}_i^{\hat{f}}|+\frac{B_{\tilde{w}}}{n^+n^-}\sup_{d_\infty(f,\hat{f})\leq\eta_K}\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}(|\ell_{ij}^f-\ell_{ij}^{\hat{f}}|)$$

$$\leqslant \frac{B_\ell\varphi_{\tilde{w}}}{n^+}\sup_{d_\infty(f,\hat{f})\leq\eta_K}\sum_{i=1}^{n^+}|f(x_i^+)-\hat{f}(x_i^+)|$$

$$+\frac{B_{\tilde{w}}\varphi_\ell}{n^+n^-}\sup_{d_\infty(f,\hat{f})\leq\eta_K}\sum_{i=1}^{n^+}\sum_{j=1}^{n^-}(|f(x_i^+)-\hat{f}(x_i^+)-f(x_j^-)+\hat{f}(x_j^-)|)$$

$$\leqslant (2B_{\tilde{w}}\varphi_\ell+B_\ell\varphi_{\tilde{w}})\cdot d_\infty(f,\hat{f})\leqslant(2B_{\tilde{w}}\varphi_\ell+B_\ell\varphi_{\tilde{w}})\eta_K \tag{附 6}$$

结合式(附 3)、式(附 4)、式(附 5)、式(附 6)可以得到:

$$\mathfrak{R}_S^p((\alpha\circ\mathcal{F})\cdot(\ell\circ\mathcal{F}))\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta_K+6C\sum_{k=1}^{K}\eta_k\sqrt{\log(\mathcal{N}(\eta_k,\mathcal{F},d_\infty))}$$

$$=(4B_{\tilde{w}}\phi_\ell+2B_\ell\phi_{\tilde{w}})\eta_{K+1}+12C\sum_{k=1}^{K}(\eta_k-\eta_{k+1})\sqrt{\log(\mathcal{N}(\eta_k,\mathcal{F},d_\infty))}$$

$$\leqslant(4B_{\tilde{w}}\phi_\ell+2B_\ell\phi_{\tilde{w}})\eta_{K+1}+12C\int_{\eta_{K+1}}^{\eta_0}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d_\infty)}\,\mathrm{d}\epsilon$$

$$\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta+\frac{12C_0}{\sqrt{n}}\int_{\eta}^{B_\ell}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d_\infty)}\,\mathrm{d}\epsilon,$$

其中,$C_0=\dfrac{2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}}}{2}\sqrt{2+\rho+1/\rho}$。在 $\log(\mathcal{N}(\epsilon,\mathcal{F},d_\infty))\leqslant A\log\left(\dfrac{b}{\epsilon}\right)$ 的假设下可以得到:

$$\mathfrak{R}_S^p((\alpha\circ\mathcal{F})\cdot(\ell\circ\mathcal{F}))\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta+\frac{12C_0}{\sqrt{n}}\int_{\eta}^{B_\ell}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F},d_\infty)}\,\mathrm{d}\epsilon$$

$$\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta+\frac{12C_0}{\sqrt{n}}\int_{\eta}^{B_\ell}\sqrt{A\log\left(\frac{B_\ell}{\epsilon}\right)}\,\mathrm{d}\epsilon$$

$$\leqslant(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})\eta+\frac{12C_0}{\sqrt{n}}\cdot B_\ell\sqrt{A\log\left(\frac{B_\ell}{\eta}\right)}\ 。$$

取 $\eta=\dfrac{1}{\sqrt{n}}$,可以得到结论(2)。

证毕。

引理 10. 假设正负样本数量的比率 $\rho=n^+/n^-$ 是固定的,$\tilde{w}\in[1,B_{\tilde{w}}]$,$\ell\in[0,B_\ell]$。至少以 $1-\delta$ 的概率有:

$$\mathfrak{R}_{n^+,n^-}^p(\mathcal{G})\leqslant\mathfrak{R}_S^p(\mathcal{G})+B_{\tilde{w}}B_\ell\sqrt{\frac{2(2+\rho+1/\rho)}{n}\log\frac{1}{\delta}},$$

其中,

$$\mathcal{G}=\{g(x_1,x_2;f,\tilde{w},\ell)=\tilde{w}(\ell(f(\boldsymbol{x}_1)-\mu^+)/B_\ell)\cdot\ell(f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2))\mid f\in\mathcal{F}\}。$$

证明. 考虑两个仅有一个样本不同的数据集 $s=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ 和 $s'=\{(x_i,y_i)\}_{i=1}^{k-1}\cup(x'_k,y_k)\cup\{(x_i,y_i)\}_{i=k+1}^n$。

一方面,如果 $y_k=1$,则可以得到:

$$|\mathfrak{R}_S^p(\mathcal{G})-\mathfrak{R}_{S'}^p(\mathcal{G})|\leqslant\mathbb{E}_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{n^-}\frac{\sigma_k^++\sigma_j^-}{2n^+n^-}(\tilde{w}_k\ell(f(\boldsymbol{x}_k)-f(\boldsymbol{x}_j^-))-\tilde{w}'_k\ell(f(\boldsymbol{x}'_k)-f(\boldsymbol{x}_j^-)))\right|$$

$$\leqslant\mathbb{E}_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\sum_{j=1}^{n^-}\left[\frac{|\sigma_k^++\sigma_j^-|}{2n^+n^-}\cdot|\tilde{w}_k\ell(f(\boldsymbol{x}_k)-f(\boldsymbol{x}_j^-))-\tilde{w}'_k\ell(f(\boldsymbol{x}'_k)-f(\boldsymbol{x}_j^-))|\right]$$

$$\leqslant \frac{1}{n^+ n^-} \cdot \sup_{f\in\mathcal{F}} \sum_{j=1}^{n^-} \left| \widetilde{w}_k \ell(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_j^-)) - \widetilde{w}'_k \ell(f(\boldsymbol{x}'_k) - f(\boldsymbol{x}_j^-)) \right|$$

$$\leqslant \frac{1}{n^+ n^-} \cdot \sum_{j=1}^{n^-} 2B_{\widetilde{w}} B_\ell = \frac{2B_{\widetilde{w}} B_\ell}{n^+}。$$

另一方面,如果 $y_k = -1$,类似地可以得到:

$$\left| \Re_{\mathcal{S}}^{\rho}(\mathcal{G}) - \Re_{\mathcal{S}'}^{\rho}(\mathcal{G}) \right| \leqslant \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n^+} \frac{\sigma_k^+ + \sigma_j^-}{2n^+ n^-} (\widetilde{w}_i \ell(f(\boldsymbol{x}_i^+) - f(\boldsymbol{x}_k)) - \widetilde{w}_i \ell(f(\boldsymbol{x}_i^+) - f(\boldsymbol{x}'_k))) \right|$$

$$\leqslant \frac{1}{n^+ n^-} \cdot \sup_{f\in\mathcal{F}} \sum_{i=1}^{n^+} \left| \widetilde{w}_i \ell(f(\boldsymbol{x}_i^+) - f(\boldsymbol{x}_k)) - \widetilde{w}_i \ell(f(\boldsymbol{x}_i^+) - f(\boldsymbol{x}'_k)) \right| \leqslant \frac{2B_{\widetilde{w}} B_\ell}{n^-}。$$

因此,根据 McDiarmid 不等式(引理 7),对于任意 $\delta \in (0,1)$,以至少 $1-\delta$ 的概率,都有

$$\Re_{n^+,n^-}^{\rho}(\mathcal{G}) \leqslant \Re_{\mathcal{S}}^{\rho}(\mathcal{G}) + \sqrt{\frac{1}{2} \cdot \left[n^+ \left(\frac{2B_{\widetilde{w}} B_\ell}{n^+} \right)^2 + n^- \left(\frac{2B_{\widetilde{w}} B_\ell}{n^-} \right)^2 \right] \cdot \log \frac{1}{\delta}}$$

$$= \Re_{\mathcal{S}}^{\rho}(\mathcal{G}) + B_{\widetilde{w}} B_\ell \sqrt{2\left(\frac{1}{n^+} + \frac{1}{n^-} \right) \log \frac{1}{\delta}} = \Re_{\mathcal{S}}^{\rho}(\mathcal{G}) + B_{\widetilde{w}} B_\ell \sqrt{2(2+\rho+1/\rho) \log \frac{1}{\delta} \cdot \frac{1}{n}}。$$

证毕。

定义 5. 一个随机过程 $\theta \mapsto X_\theta$,其索引集为 $\mathcal{T}$,如果对所有 $\theta, \theta' \in \mathcal{T}$ 和所有 $\lambda \in \mathbb{R}$,满足以下条件,则称其对(伪)度量 d 是 sub-Gaussian 的:

$$\mathbb{E}[\exp(\lambda \cdot (X_\theta - X_{\theta'}))] \leqslant \left(\frac{\lambda^2 d(\theta,\theta')^2}{2} \right)。$$

记

$$T_f(\boldsymbol{\sigma}) = \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \frac{\sigma_i^+ + \sigma_j^-}{2n^+ n^-} \cdot \widetilde{w}_i \ell(f(\boldsymbol{x}_i^+) - f(\boldsymbol{x}_j^-)),$$

其中,f 从索引集 $\mathcal{T}$ 中选取,回顾经验成对 Rademacher 复杂度:

$$\Re_{\mathcal{S}}^{\rho}((\widetilde{w} \circ \mathcal{F}) \cdot (\ell \circ \mathcal{F})) = \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{f\in\mathcal{F}} |T_f(\boldsymbol{\sigma})|]。$$

在以下引理中,我们首先证明对于 Rademacher 随机变量 $\boldsymbol{\sigma} C T_f(\boldsymbol{\sigma})$ 满足 sub-Gaussian 性质。

引理 11. 成对 Rademacher 复杂度的 sub-Gaussian 性质. 对于一个假设空间 $\mathcal{F}$,两个函数 $f, \widetilde{f} \in \mathcal{F}$,以及一个损失函数 ℓ,假设 ℓ 是 ϕ_ℓ- Lipschitz 连续的,加权函数 $\widetilde{w}$ 是 $\phi_{\widetilde{w}}$- Lipschitz 连续的。那么对于所有 $\lambda \in \mathbb{R}$,可以得到:

$$\mathbb{E}[\exp(\lambda \cdot (T_f(\boldsymbol{\sigma}) - T_{\widetilde{f}}(\boldsymbol{\sigma})))] \leqslant \exp\left(\frac{C^2\lambda^2}{2} \max_{x\in\mathcal{S}} |f(\boldsymbol{x}) - \widetilde{f}(\boldsymbol{x})|^2 \right),$$

其中,$C = \dfrac{2B_{\widetilde{w}}\phi_\ell + B_\ell \phi_{\widetilde{w}}}{2} \sqrt{\dfrac{1}{n^+} + \dfrac{1}{n^-}}$。

证明. 首先证明 $T_f(\boldsymbol{\sigma})$ 关于 $\boldsymbol{\sigma}$ 满足有界差分条件。考虑两个 Rademacher 随机变量 $\boldsymbol{\sigma}$ 和 $\boldsymbol{\sigma}'$,其中只有一个值不同。

当不同的值对应的样本为正例时,记:

$$\boldsymbol{\sigma} = (\sigma_1^+, \cdots, \sigma_k^+, \cdots, \sigma_{n^+}^+, \sigma_1^-, \cdots, \sigma_{n^-}^-),$$

$$\boldsymbol{\sigma}' = (\sigma_1^+, \cdots, \sigma_k'^+, \cdots, \sigma_{n^+}^+, \sigma_1^-, \cdots, \sigma_{n^-}^-),$$

$$\ell_{ij}^f = \ell(f(\boldsymbol{x}_i^+) - f(\boldsymbol{x}_j^-)),$$

$$\widetilde{w}_i^f = \widetilde{w}(f(\boldsymbol{x}_i^+) - \mathbb{E}_{x\in\mathcal{S}^+}[f(\boldsymbol{x})]),$$

那么可以得到:

$$\left| (T_f(\boldsymbol{\sigma}) - T_{\widehat{f}}(\boldsymbol{\sigma})) - (T_f(\boldsymbol{\sigma}') - T_{\widehat{f}}(\boldsymbol{\sigma}')) \right|$$

$$= \frac{1}{n^+ n^-} \left| \sum_{j=1}^{n^-} \frac{\sigma_k^+ - \sigma_k'^+}{2} \cdot \widetilde{w}_k^f \ell_{kj}^f - \sum_{j=1}^{n^-} \frac{\sigma_k^+ - \sigma_k'^+}{2} \cdot \widetilde{w}_k^{\widetilde{f}} \ell_{kj}^{\widetilde{f}} \right|$$

$$\leqslant \frac{1}{n^+n^-}\left|\frac{\sigma_k^+-\sigma'^+_k}{2}\right|\cdot\left|\sum_{j=1}^{n^-}(\tilde{w}_k^f\ell_{kj}^f-\tilde{w}_k^{\tilde{f}}\ell_{kj}^{\tilde{f}})\right|$$

$$\leqslant \frac{1}{n^+n^-}\left|\sum_{j=1}^{n^-}(\tilde{w}_k^f\ell_{kj}^f-\tilde{w}_k^f\ell_{kj}^{\tilde{f}}+\tilde{w}_k^f\ell_{kj}^{\tilde{f}}-\tilde{w}_k^{\tilde{f}}\ell_{kj}^{\tilde{f}})\right|$$

$$\leqslant \frac{B_{\tilde{w}}}{n^+n^-}\sum_{j=1}^{n^-}|\ell_{kj}^f-\ell_{kj}^{\tilde{f}}|+\frac{B_\ell}{n^+}|\tilde{w}_k^f-\tilde{w}_k^{\tilde{f}}|$$

$$\leqslant \frac{B_{\tilde{w}}\varphi_\ell}{n^+n^-}\sum_{j=1}^{n^-}|f(\boldsymbol{x}_k^+)-\tilde{f}(\boldsymbol{x}_k^+)-f(\boldsymbol{x}_j^-)+\tilde{f}(\boldsymbol{x}_j^-)|+\frac{B_\ell\varphi_{\tilde{w}}}{n^+}|f(\boldsymbol{x}_k^+)-\tilde{f}(\boldsymbol{x}_k^+)|$$

$$\leqslant \frac{B_{\tilde{w}}\varphi_\ell}{n^+}\max_{x\in\mathcal{S}^+}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|+\frac{B_{\tilde{w}}\varphi_\ell}{n^+}\max_{x\in\mathcal{S}^-}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|+\frac{B_\ell\varphi_{\tilde{w}}}{n^+}\max_{x\in\mathcal{S}^+}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|$$

$$\leqslant \frac{2B_{\tilde{w}}\varphi_\ell+B_\ell\varphi_{\tilde{w}}}{n^+}\max_{x\in\mathcal{S}}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|\triangleq c_k^+。$$

当不同的值对应的样本为负例时，记

$$\boldsymbol{\sigma}=(\sigma^+,\cdots,\sigma_{n^+}^+,\sigma_1^-,\cdots,\sigma_k^-,\cdots,\sigma_{n^-}^-),$$

$$\boldsymbol{\sigma}'=(\sigma^+,\cdots,\sigma_{n^+}^+,\sigma_1^-,\cdots,\sigma'^-_k,\cdots,\sigma_{n^-}^-),$$

那么可以得到：

$$|(T_f(\boldsymbol{\sigma})-T_{\hat{f}}(\boldsymbol{\sigma}))-(T_f(\boldsymbol{\sigma}')-T_{\hat{f}}(\boldsymbol{\sigma}'))|$$

$$\leqslant \frac{1}{n^+n^-}\left|\sum_{i=1}^{n^+}(\tilde{w}f_i\ell_{ik}^f-\tilde{w}_i^f\ell_{ik}^{\tilde{f}}+\tilde{w}_i^f\ell_{ik}^{\tilde{f}}-\tilde{w}_i^{\tilde{f}}\ell_{ik}^{\tilde{f}})\right|$$

$$\leqslant \frac{B_{\tilde{w}}}{n^+n^-}\sum_{i=1}^{n^+}|\ell_{ik}^f-\ell_{ik}^{\tilde{f}}|+\frac{B_\ell}{n^+n^-}\sum_{i=1}^{n^+}|\tilde{w}f_i-\tilde{w}_i^{\tilde{f}}|$$

$$\leqslant \frac{B_{\tilde{w}}\phi_\ell}{n^+n^-}\sum_{i=1}^{n^+}|f(\boldsymbol{x}_i^+)-\tilde{f}(\boldsymbol{x}_i^+)-f(\boldsymbol{x}_k^-)+\tilde{f}(\boldsymbol{x}_k^-)|+\frac{B_\ell\phi_{\tilde{w}}}{n^+n^-}\sum_{i=1}^{n^+}|f(\boldsymbol{x}_i^+)-\tilde{f}(\boldsymbol{x}_i^+)|$$

$$\leqslant \frac{2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}}}{n^-}\max_{x\in\mathcal{S}}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|\triangleq c_k^-。$$

因此，根据引理 4，我们可以选择：

$$v=\frac{1}{4}\left(\sum_{k=1}^{n^+}(c_k^+)^2+\sum_{k=1}^{n^-}(c_k^-)^2\right)$$

$$=\frac{(2B_{\tilde{w}}\phi_\ell+B_\ell\phi_{\tilde{w}})^2}{4}\left(\frac{1}{n^+}+\frac{1}{n^-}\right)\cdot\max_{x\in\mathcal{S}}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|^2=C^2\max_{x\in\mathcal{S}}|f(\boldsymbol{x})-\tilde{f}(\boldsymbol{x})|^2,$$

使得结论成立。

证毕。

附录 C. 收敛性分析

引理 12（文献[58]中的引理 1）. 对于算法 1 中的更新规则，设

$$\boldsymbol{p}_k=\begin{cases}0, k=1,\\ \dfrac{\beta}{1-\beta}(\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}+s\alpha\boldsymbol{g}_{k-1}), k\geqslant 2,\end{cases}$$

$$\boldsymbol{z}_k=\boldsymbol{\theta}_k+p_k,$$

那么，对于任意 $k\geqslant 1$，可以得到：

$$\boldsymbol{z}_{k+1}=\boldsymbol{z}_k-\frac{\alpha}{1-\beta}\boldsymbol{g}_k。$$

引理 13（文献[58]中的引理 4）. 设假设 1 和假设 2 成立，那么对于所有 $k\geqslant 1$，可以得到：

$$\mathbb{E}\left[\nabla e(\boldsymbol{z}_k;\mu_k^+)-\nabla e(\boldsymbol{\theta}_k;\mu_k^+)^2\right]\leqslant\frac{L_u^2\beta^2\alpha^2(G^2+\kappa_e^2)}{(1-\beta)^2}。$$

引理 14. 设 $\boldsymbol{\theta}_k$ 按照算法 1 中生成,那么可以得到 $\boldsymbol{\theta}$ 差的上界如下:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k^2\right]\leqslant\alpha^2(G^2+\kappa_e^2)\cdot\left[T(s(1-\beta)-1)^2\beta^2/(1-\beta^2)+2(1+s\beta)^2\right]。$$

证明. 根据 $\boldsymbol{\theta}$ 更新规则,对于 $k\geqslant 2$,可以得到:

$$\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k=-\alpha\boldsymbol{g}_k+\beta(\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1})-s\alpha\beta(\boldsymbol{g}_k-\boldsymbol{g}_{k-1})=$$

$$\beta(\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1})-\alpha(1+s\beta)\boldsymbol{g}_k+s\alpha\beta\boldsymbol{g}_{k-1}=-\alpha(1+s\beta)\sum_{t=1}^{k}\beta^{k-t}\boldsymbol{g}_t+s\alpha\sum_{t=1}^{k-1}\beta^{k-t}\boldsymbol{g}_t+\beta(\boldsymbol{\theta}_1-\gamma_1^s)$$

$$=\alpha(s(1-\beta)-1)\sum_{t=1}^{k-1}\beta^{k-t}\boldsymbol{g}_t-\alpha(1+s\beta)\boldsymbol{g}_k+\beta(\boldsymbol{\theta}_1-\gamma_1^s)。$$

通过设置 $\gamma_1^s=\boldsymbol{\theta}_1$,可以得到:

$$\mathbb{E}\left[\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k^2\right]=\mathbb{E}\left[\alpha(s(1-\beta)-1)\sum_{t=1}^{k-1}\beta^{k-t}\boldsymbol{g}_t-\alpha(1+s\beta)\boldsymbol{g}_k^2\right]$$

$$\leqslant 2\mathbb{E}\left[\alpha(s(1-\beta)-1)\sum_{t=1}^{k-1}\beta^{k-t}\boldsymbol{g}_t^2\right]+2\mathbb{E}\left[\alpha(1+s\beta)\boldsymbol{g}_k^2\right]$$

$$\leqslant 2(k-1)\cdot\alpha^2(s(1-\beta)-1)^2\sum_{t=1}^{k-1}\beta^{2(k-t)}\mathbb{E}\left[\boldsymbol{g}_t^2\right]+2\alpha^2(1+s\beta)^2\mathbb{E}\left[\boldsymbol{g}_k^2\right]$$

$$\leqslant 2(k-1)(G^2+\kappa_e^2)\cdot\alpha^2(s(1-\beta)-1)^2\sum_{t=1}^{k-1}\beta^{2(k-t)}+2(G^2+\kappa_e^2)\cdot\alpha^2(1+s\beta)^2$$

$$\leqslant 2\alpha^2(G^2+\kappa_e^2)\left[\frac{(k-1)(s(1-\beta)-1)^2\beta^2}{1-\beta^2}+(1+s\beta)^2\right]。$$

通过取 $k=1,2,\cdots,T$ 的平均值,证明成立。

证毕。

引理 15. 设假设 1、假设 2 和假设 3 成立。那么对于所有 $k\geqslant 1$,μ_k^+ 的估计偏差的上界如下:

$$\mathbb{E}\left[\nabla e(\boldsymbol{\theta}_k;\mu_k^+)-\nabla e(\boldsymbol{\theta}_k;\hat{u}_k)^2\right]\leqslant 6\alpha^2L_u^2\phi_f^2(G^2+\kappa_e^2)\lambda^{-1}\cdot\left[(s(1-\beta)-1)^2\beta^2T/(1-\beta^2)+2(1+s\beta)^2\right]+3L_u^2\lambda\kappa_\mu^2$$

证明. 根据 $\nabla e(\boldsymbol{\theta}_k;\cdot)$ 的平滑性,可以得到:

$$\nabla e(\boldsymbol{\theta}_k;\mu_k^+)-\nabla e(\boldsymbol{\theta}_k;\hat{u}_k)^2\leqslant L_u^2\mu_k^+-\hat{u}_k^2.\tag{附 7}$$

记 $\zeta_{1:k}=\{\zeta_t\}_{t=1}^k$。通过代入 $\hat{\mu}_k$ 的更新规则(式(6)),可以得到如下的递归式:

$$\hat{u}_k-\mu_k^{+2}=(1-\lambda)\hat{u}_{k-1}+\lambda\mu_k^+(\zeta_k)+(1-\lambda)(\mu_k^+(\zeta_k)-\mu_{k-1}^+(\zeta_k))-\mu_k^{+2}$$

$$=(1-\lambda)(\hat{u}_{k-1}-\mu_{k-1}^+)-\mu_k^++\mu_k^+(\zeta_k)+(1-\lambda)(\mu_{k-1}^+-\mu_{k-1}^+(\zeta_k))^2$$

$$=(1-\lambda)^2\hat{u}_{k-1}-\mu_{k-1}^{+2}+-\mu_k^++\mu_k^+(\zeta_k)+(1-\lambda)(\mu_{k-1}^+-\mu_{k-1}^+(\zeta_k))^2+$$

$$(\hat{u}_{k-1}-\mu_{k-1}^+)\cdot\left[-\mu_k^++\mu_k^+(\zeta_k)+(1-\lambda)(\mu_{k-1}^+-\mu_{k-1}^+(\zeta_k))\right]。$$

根据假设 3,$\mathbb{E}_{\zeta_{1:k-1}}\left[-\mu_k^++\mu_k^+(\zeta_k)\right]=0$,因此可以得到:

$$\mathbb{E}_{\zeta_{1:k-1}}\left[\hat{u}_k-\mu_k^{+2}\right]=$$

$$(1-\lambda)^2\mathbb{E}_{\zeta_{1:k-1}}\left[\hat{u}_{k-1}-\mu_{k-1}^{+2}\right]+\mathbb{E}_{\zeta_{1:k-1}}\left[-\mu_k^++\mu_k^+(\zeta_k)+(1-\lambda)(\mu_{k-1}^+-\mu_{k-1}^+(\zeta_k))^2\right]$$

$$=(1-\lambda)^2\mathbb{E}_{\zeta_{1:k-1}}\left[\hat{u}_{k-1}-\mu_{k-1}^{+2}\right]+\mathbb{E}_{\zeta_{1:k-1}}\left[-\mu_k^++\mu_{k-1}^++\mu_k^+(\zeta_k)-\mu_{k-1}^+(\zeta_k)+\lambda(\mu_{k-1}^+(\zeta_k)-\mu_{k-1}^+)^2\right]$$

$$\leqslant(1-\lambda)\mathbb{E}_{\zeta_{1:k-1}}\left[\hat{u}_{k-1}-\mu_{k-1}^{+2}\right]+3\mathbb{E}_{\zeta_{1:k}}\left[\mu_k^+-\mu_{k-1}^{+2}+\mu_k^+(\zeta_k)-\mu_{k-1}^+(\zeta_k)^2+\lambda^2\mu_{k-1}^+(\zeta_k)-\mu_{k-1}^{+2}\right]。$$

根据假设 1,可以得到 $\mu_k^+-\mu_{k-1}^{+2}\leqslant\phi_f^2\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}^2$,$\mu_k^+(\zeta_k)-\mu_{k-1}^+(\zeta_k)^2\leqslant\phi_f^2\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}^2$。进而根据假设 2,可以得到 $\mathbb{E}\left[\mu_{k-1}^+(\zeta_k)-\mu_{k-1}^{+2}\right]\leqslant\kappa_\mu^2$,可以推出:

$$\mathbb{E}\left[\hat{u}_k-\mu_k^{+2}\right]\leqslant(1-\lambda)\mathbb{E}\left[\hat{u}_{k-1}-\mu_{k-1}^{+2}\right]+6\phi_f^2\mathbb{E}\left[\boldsymbol{\theta}_k-\boldsymbol{\theta}_{k-1}^2\right]+3\lambda^2\kappa_\mu^2。$$

通过对 $k=2,\cdots,T+1$ 求和,上述结果可以重新组织为

$$\lambda\sum_{k=1}^{T}\mathbb{E}\left[\hat{u}_k-\mu_k^{+2}\right]\leqslant\mathbb{E}\left[\hat{u}_1-\mu_1^{+2}\right]-\mathbb{E}\left[\hat{u}_{T+1}-\mu_{T+1}^{+2}\right]+6\phi_f^2\,\mathbb{E}\left[\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k{}^2\right]+3T\lambda^2\kappa_\mu^2,$$

如果 $\hat{u}_1$ 初始化为 μ_1^+,可以得到:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\hat{u}_k-\mu_k^{+2}\right]\leqslant\frac{6\phi_f^2}{\lambda T}\sum_{k=1}^{T}\mathbb{E}\left[\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k{}^2\right]+3\lambda\kappa_\mu^2$$

$$\overset{\text{引理14}}{\leqslant}6\alpha^2\phi_f^2(G^2+\kappa_e^2)\lambda^{-1}\cdot\left[(s(1-\beta)-1)^2\beta^2T/(1-\beta^2)+2(1+s\beta)^2\right]+3\lambda\kappa_\mu^2。$$

将上述不等式代入式(附 7)中即可完成该证明。

证毕。

引理 16. 设假设 1、假设 2 和假设 3 成立。那么对于所有 $k\geqslant1$,可以得到:

$$\mathbb{E}\left[e(\boldsymbol{z}_{k+1})-e(\boldsymbol{z}_k)\right]$$

$$\leqslant\frac{L_u^2(G^2+\kappa_e^2)}{2(1-\beta)^3}\cdot\alpha^3(1+3L_u\beta^2)-\frac{\alpha}{2(1-\beta)}\mathbb{E}\left[\left\|\nabla e(\boldsymbol{\theta}_k)\right\|^2\right]+\frac{3\alpha}{2(1-\beta)}\mathbb{E}\left[\left\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\right\|^2\right].$$

证明. 为了简化表达,将 $e(\boldsymbol{\theta}_k;\mu_k^+)$ 简记为 $e(\boldsymbol{\theta}_k)$。根据假设 1,$e(\cdot)$ 是 L- 平滑的,那么可以得到:

$$e(\boldsymbol{z}_{k+1})-e(\boldsymbol{z}_k)\leqslant\nabla e(\boldsymbol{z}_k)^\top(\boldsymbol{z}_{k+1}-\boldsymbol{z}_k)+\frac{L_\theta}{2}\left\|\boldsymbol{z}_{k+1}-\boldsymbol{z}_k\right\|^2=\nabla e(\boldsymbol{z}_k)^\top(\boldsymbol{z}_{k+1}-\boldsymbol{z}_k)+\frac{L_\theta\alpha^2}{2(1-\beta)^2}\left\|\boldsymbol{g}_k\right\|^2$$

(附 8)

根据引理 12,可以得到:

$$\nabla e(\boldsymbol{z}_k)^\top(\boldsymbol{z}_{k+1}-\boldsymbol{z}_k)\leqslant-\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^\top\boldsymbol{g}_k\leqslant-\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^\top\nabla e(\boldsymbol{\theta}_k)+\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^\top(\nabla e(\boldsymbol{\theta}_k)-\boldsymbol{g}_k)$$

(附 9)

然后,进一步分解第一项可以得到:

$$-\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^\top\nabla e(\boldsymbol{\theta}_k)=-\frac{\alpha}{1-\beta}\left\|\nabla e(\boldsymbol{\theta}_k)\right\|^2-\frac{\alpha}{1-\beta}(\nabla e(\boldsymbol{z}_k)^\top-\nabla e(\boldsymbol{\theta}_k)^\top)\nabla e(\boldsymbol{\theta}_k)\quad\text{(附 10)}$$

根据 Cauchy-Schwarz 不等式,可以得到:

$$-\frac{\alpha}{1-\beta}(\nabla e(\boldsymbol{z}_k)^\top-\nabla e(\boldsymbol{\theta}_k)^\top)\nabla e(\boldsymbol{\theta}_k)=-\frac{\alpha}{1-\beta}\cdot\sqrt{2}(\nabla e(\boldsymbol{z}_k)^\top-\nabla e(\boldsymbol{\theta}_k)^\top)\cdot\frac{1}{\sqrt{2}}\nabla e(\boldsymbol{\theta}_k)$$

$$\leqslant\frac{\alpha}{1-\beta}\left\|\nabla e(\boldsymbol{z}_k)-\nabla e(\boldsymbol{\theta}_k)\right\|^2+\frac{\alpha}{4(1-\beta)}\left\|\nabla e(\boldsymbol{\theta}_k)\right\|^2\quad\text{(附 11)}$$

类似地,可以将分解第二项得到:

$$\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^\top(\nabla e(\boldsymbol{\theta}_k)-\boldsymbol{g}_k)=\frac{\alpha}{1-\beta}(\nabla e(\boldsymbol{z}_k)^\top-\nabla e(\boldsymbol{\theta}_k)^\top)(\nabla e(\boldsymbol{\theta}_k)-\nabla e(\boldsymbol{\theta}_k;\hat{u}_k))$$

$$+\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{\theta}_k)^\top(\nabla e(\boldsymbol{\theta}_k)-\nabla e(\boldsymbol{\theta}_k;\hat{u}_k))$$

$$+\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^\top(\nabla e(\boldsymbol{\theta}_k;\hat{u}_k)-\boldsymbol{g}_k)\quad\text{(附 12)}$$

其中,

$$\frac{\alpha}{1-\beta}(\nabla e(\boldsymbol{z}_k)^{\top}-\nabla e(\boldsymbol{\theta}_k)^{\top})(\nabla e(\boldsymbol{\theta}_k)-\nabla e(\boldsymbol{\theta}_k;\hat{u}_k))$$

$$\leqslant\frac{\alpha}{2(1-\beta)}\|\nabla e(\boldsymbol{z}_k)-\nabla e(\boldsymbol{\theta}_k)\|^2+\frac{\alpha}{2(1-\beta)}\|\nabla e(\boldsymbol{\theta}_k)-\nabla e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2,$$

$$\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{\theta}_k)^{\top}(\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k))\leqslant\frac{\alpha}{4(1-\beta)}\|\nabla e(\boldsymbol{\theta}_k)\|^2+\frac{\alpha}{1-\beta}\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2$$

(附 13)

通过将式(附 9)～式(附 13)带入到式(附 8)中,可以得到:

$$e(\boldsymbol{z}_{k+1})-e(\boldsymbol{z}_k)\leqslant\frac{3\alpha}{2(1-\beta)}\|\nabla e(\boldsymbol{z}_k)-\nabla e(\boldsymbol{\theta}_k)\|^2-\frac{\alpha}{2(1-\beta)}\|\nabla e(\boldsymbol{\theta}_k)\|^2+\frac{3\alpha}{2(1-\beta)}\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2$$

$$+\frac{\alpha}{1-\beta}\nabla e(\boldsymbol{z}_k)^{\top}(\nabla e(\boldsymbol{\theta}_k;\hat{u}_k)-\boldsymbol{g}_k)+\frac{L_\theta\alpha^2}{2(1-\beta)^2}\|\boldsymbol{g}_k\|^2。$$

两边分别求期望,可以得到:

$$\mathbb{E}[e(\boldsymbol{z}_{k+1})-e(\boldsymbol{z}_k)]$$

$$\leqslant\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{z}_k)-\nabla e(\boldsymbol{\theta}_k)\|^2]-\frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)\|^2]+\frac{L_\theta\alpha^2}{2(1-\beta)^2}\mathbb{E}[\|\boldsymbol{g}_k\|^2]$$

$$+\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2]+\frac{\alpha}{1-\beta}\mathbb{E}[\nabla e(\boldsymbol{z}_k)^{\top}(\nabla e(\boldsymbol{\theta}_k;\hat{u}_k)-\boldsymbol{g}_k)]。$$

$$\overset{\text{假设3}}{\leqslant}\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{z}_k)-\nabla e(\boldsymbol{\theta}_k)\|^2]-\frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)\|^2]$$

$$+\frac{L_\theta\alpha^2}{2(1-\beta)^2}\mathbb{E}[\|\boldsymbol{g}_k\|^2]+\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2]。$$

$$\overset{\text{引理13}}{\leqslant}\frac{3L_u^2\beta^2\alpha^3(G^2+\kappa_e^2)}{2(1-\beta)^3}-\frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)\|^2]+\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2]+\frac{L_\theta\alpha^3}{2(1-\beta)^2}\mathbb{E}[\|\boldsymbol{g}_k\|^2]。$$

$$\overset{\text{假设2}}{\leqslant}\frac{3L_u^2\beta^2\alpha^3(G^2+\kappa_e^2)}{2(1-\beta)^3}-\frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)\|^2]+\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2]$$

$$+\frac{L_\theta\alpha^2(G^2+\kappa_e^2)}{2(1-\beta)^2}=\frac{\alpha^3(L_\theta+3\beta^2L_u^2)(G^2+\kappa_e^2)}{2(1-\beta)^3}-\frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)\|^2]+\frac{3\alpha}{2(1-\beta)}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2]。$$

证毕。

定理 4. 收敛性. 记 $\boldsymbol{\theta}^*$ 是最优参数, $B'=36C_\alpha^2C_\lambda^{-1}L_u^2\phi_f^2\cdot[(s(1-\beta)-1)^2C_\beta^2+(1+s\beta)^2]$, $\Delta_e=(e(\boldsymbol{\theta}_1)-e(\boldsymbol{\theta}^*))$。在运行 T 步算法 1 的更新规则后,通过设置 $\alpha=C_\alpha/\sqrt{T}$, $\beta=\min\{1-1/\sqrt{2},1/\sqrt{3L_u}, C_\beta/\sqrt{T}\}$, $\lambda=C_\lambda/\sqrt{T}$,可以得到:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}[\|\nabla e(\boldsymbol{\theta}_k;\widetilde{w}^e)\|^2]\leqslant O(T^{-1})+B'(G^2+k_e^2)\cdot T^{-\frac{1}{2}}+2(1-\beta)C_\alpha^{-1}\Delta_e\cdot T^{-\frac{1}{2}}+9L_u^2C_\lambda k_\mu^2\cdot T^{-\frac{1}{2}}。$$

证明.根据引理 16,可以得到:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\nabla e(\boldsymbol{\theta}_k)^2\right]$$

$$\leqslant\frac{2(1-\beta)}{\alpha T}\mathbb{E}\left[e(\boldsymbol{z}_1)-e(\boldsymbol{z}_{T+1})\right]+\frac{G^2+\kappa_e^2}{(1-\beta)^2}\cdot\alpha^2(L_\theta+3L_u^2\beta^2)+\frac{3}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla e(\boldsymbol{\theta}_k)-e(\boldsymbol{\theta}_k;\hat{u}_k)\|^2\right]$$

$$\overset{\text{引理15}}{\leqslant}\frac{2(1-\beta)}{\alpha T}\mathbb{E}\left[e(\boldsymbol{z}_1)-e(\boldsymbol{z}_{T+1})\right]+\frac{G^2+\kappa_e^2}{(1-\beta)^2}\cdot\alpha^2(L_\theta+3L_u^2\beta^2)+9L_u^2\kappa_\mu^2\cdot\lambda+18\alpha^2L_u^2\phi_f^2(G^2+\kappa_e^2)\lambda^{-1}$$

$$\cdot\left[(s(1-\beta)-1)^2\beta^2T/(1-\beta^2)+2(1+s\beta)^2\right]。$$

然后,通过选择 $\beta\leqslant\min\{1-1/\sqrt{2},1/\sqrt{3L_u}\}$,上述不等式可以简化为

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla e(\boldsymbol{\theta}_k)\|^2\right]$$

$$\leqslant\frac{2(1-\beta)}{\alpha T}\mathbb{E}\left[e(\boldsymbol{z}_1)-e(\boldsymbol{z}_{T+1})\right]+2(L_u+L_\theta)(G^2+\kappa_e^2)\cdot\alpha^2+9L_u^2\kappa_\mu^2\cdot\lambda+36\alpha^2L_u^2\phi_f^2(G^2+\kappa_e^2)\lambda^{-1}$$

$$\cdot\left[(s(1-\beta)-1)^2\beta^2T+(1+s\beta)^2\right]。$$

注意到 $\boldsymbol{z}_1=\boldsymbol{\theta}_1$ 以及 $e(\boldsymbol{z}_{T+1})\geqslant e(\boldsymbol{\theta}^*)$,因此:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla e(\boldsymbol{\theta}_k)\|^2\right]$$

$$\leqslant\frac{2(1-\beta)}{\alpha T}(e(\boldsymbol{\theta}_1)-e(\boldsymbol{\theta}^*))+2(L_u+L_\theta)(G^2+\kappa_e^2)\cdot\alpha^2+9L_u^2\kappa_\mu^2\cdot\lambda+36\alpha^2L_u^2\phi_f^2(G^2+\kappa_e^2)\lambda^{-1}$$

$$\cdot\left[(s(1-\beta)-1)^2\beta^2T+(1+s\beta)^2\right]。$$

通过将 $\alpha=C_\alpha/\sqrt{T}$,$\beta=\min\{1-1/\sqrt{2},1/\sqrt{3L_u},C_\beta/\sqrt{T}\}$,$\lambda=C_\lambda/\sqrt{T}$ 带入上述结果,可以得到:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla e(\boldsymbol{\theta}_k)\|^2\right]$$

$$\leqslant\frac{2(1-\beta)}{C_\alpha\sqrt{T}}(e(\boldsymbol{\theta}_1)-e(\boldsymbol{\theta}^*))+\frac{2(L_u+L_\theta)C_\alpha^2(G^2+\kappa_e^2)}{T}+\frac{9L_u^2C_\lambda\kappa_\mu^2}{\sqrt{T}}+\frac{36C_\alpha^2L_u^2\phi_f^2(G^2+\kappa_e^2)}{C_\lambda\sqrt{T}}$$

$$\cdot\left[(s(1-\beta)-1)^2C_\beta^2+(1+s\beta)^2\right]=\frac{2(1-\beta)}{C_\alpha\sqrt{T}}(e(\boldsymbol{\theta}_1)-e(\boldsymbol{\theta}^*))+\frac{B'(G^2+\kappa_e^2)}{\sqrt{T}}+\frac{9L_u^2C_\lambda\kappa_\mu^2}{\sqrt{T}}+\frac{2(L_u+L_\theta)C_\alpha^2(G^2+\kappa_e^2)}{T}。$$

其中,$B'=36C_\alpha^2C_\lambda^{-1}L_u^2\phi_f^2\cdot\left[(s(1-\beta)-1)^2C_\beta^2+(1+s\beta)^2\right]$。

证毕。

附录 D. 实验结果可视化

D. 1. 图像检索示例

本文所提出方法的图像检索结果示例及对应 PR 曲线如图 4 所示。

图 4　在 SOP(最上方两组)、VeichleID(第 3 至第 4 组)和 iNaturalist(最下方三组)的检索结果示例(每组中上方一行由本文所提出方法预测,而下方一行来自三元组损失)

D. 2. 嵌入可视化

在三个图像检索数据集上的可视化结果如图 5、图 6 及图 7 所示。

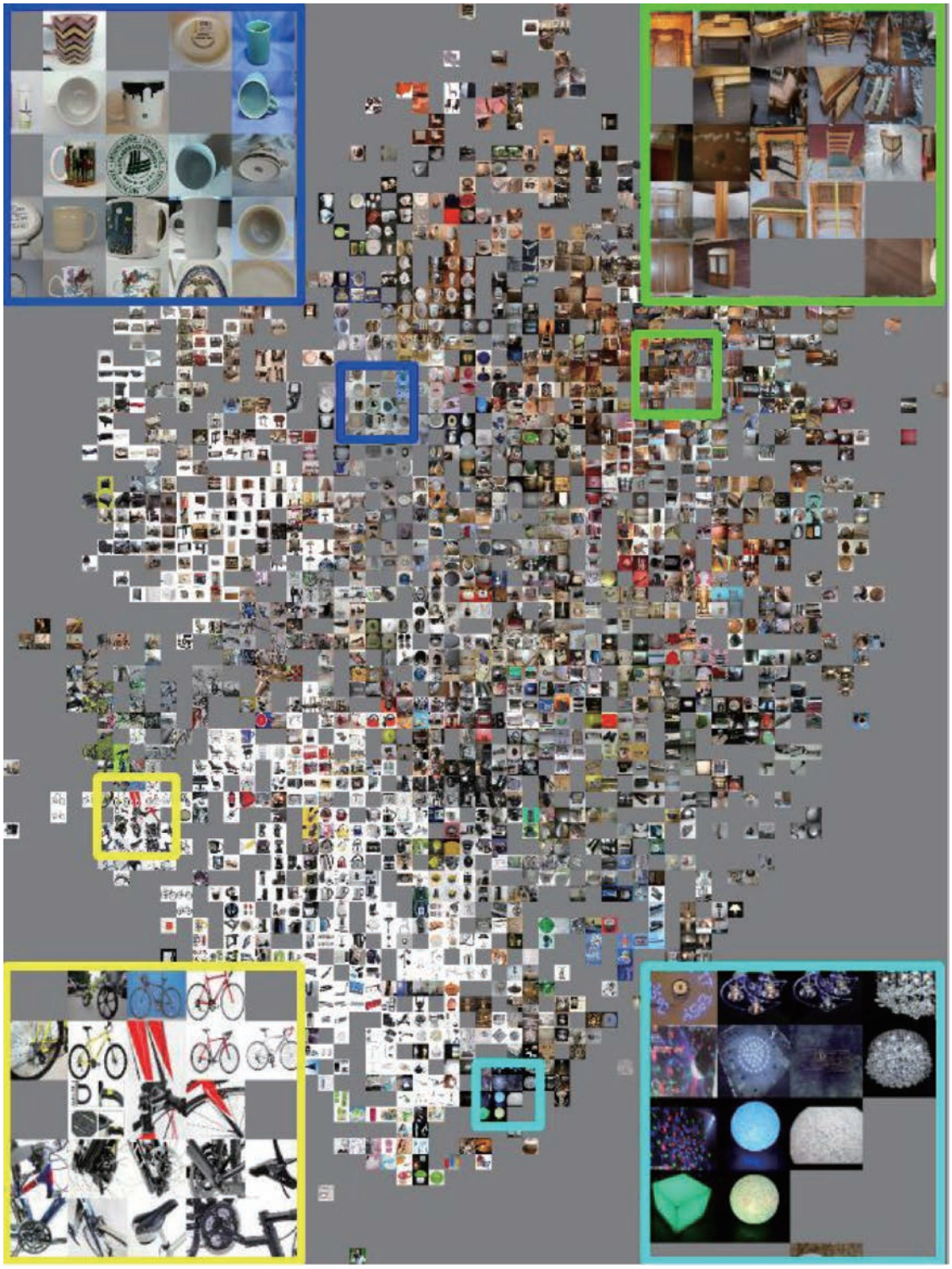


图 5 在 SOP 数据集上嵌入可视化(在显示器上观看效果最佳)

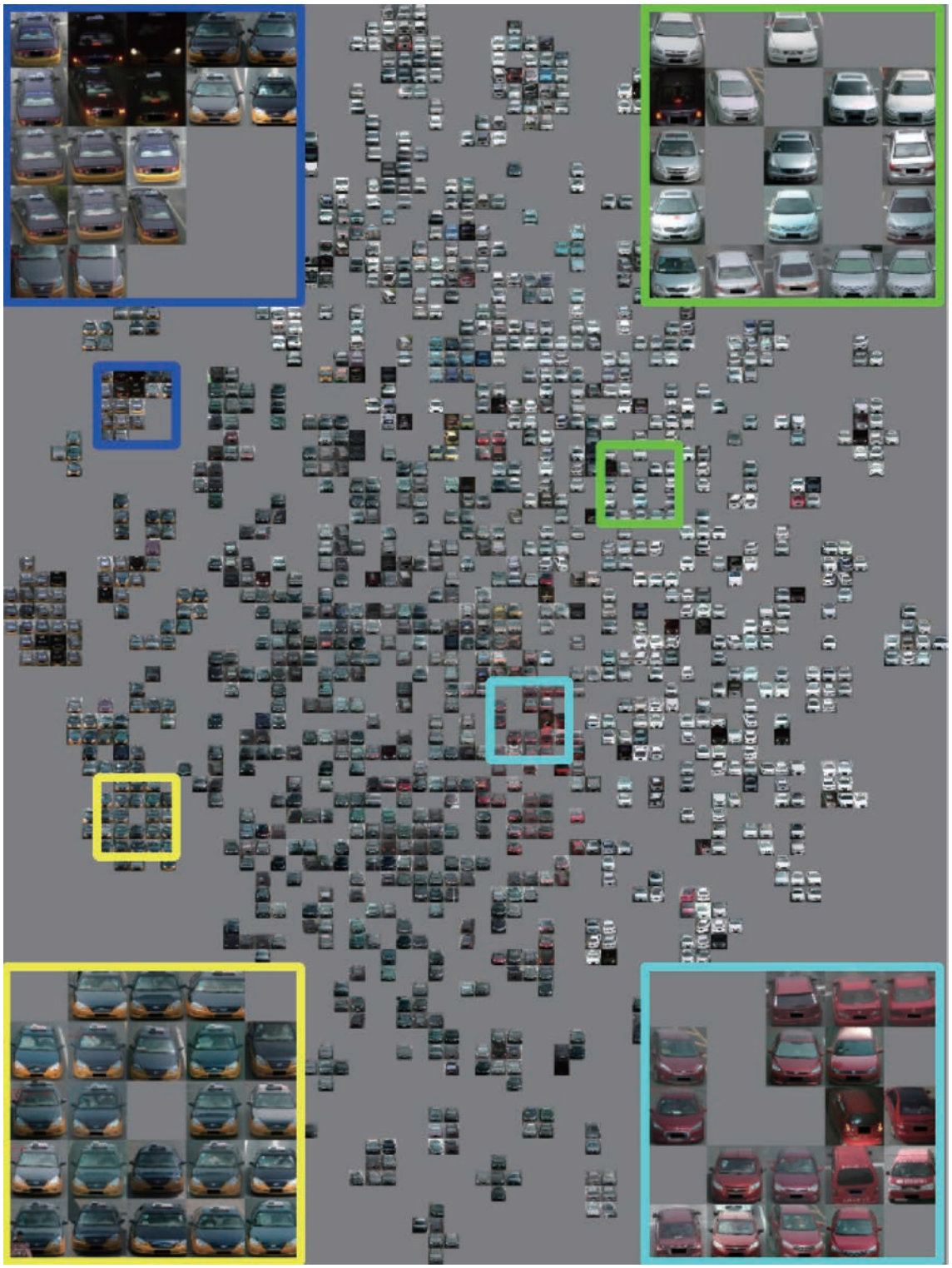


图 6　在 VehicleID 数据集上嵌入可视化(在显示器上观看效果最佳)

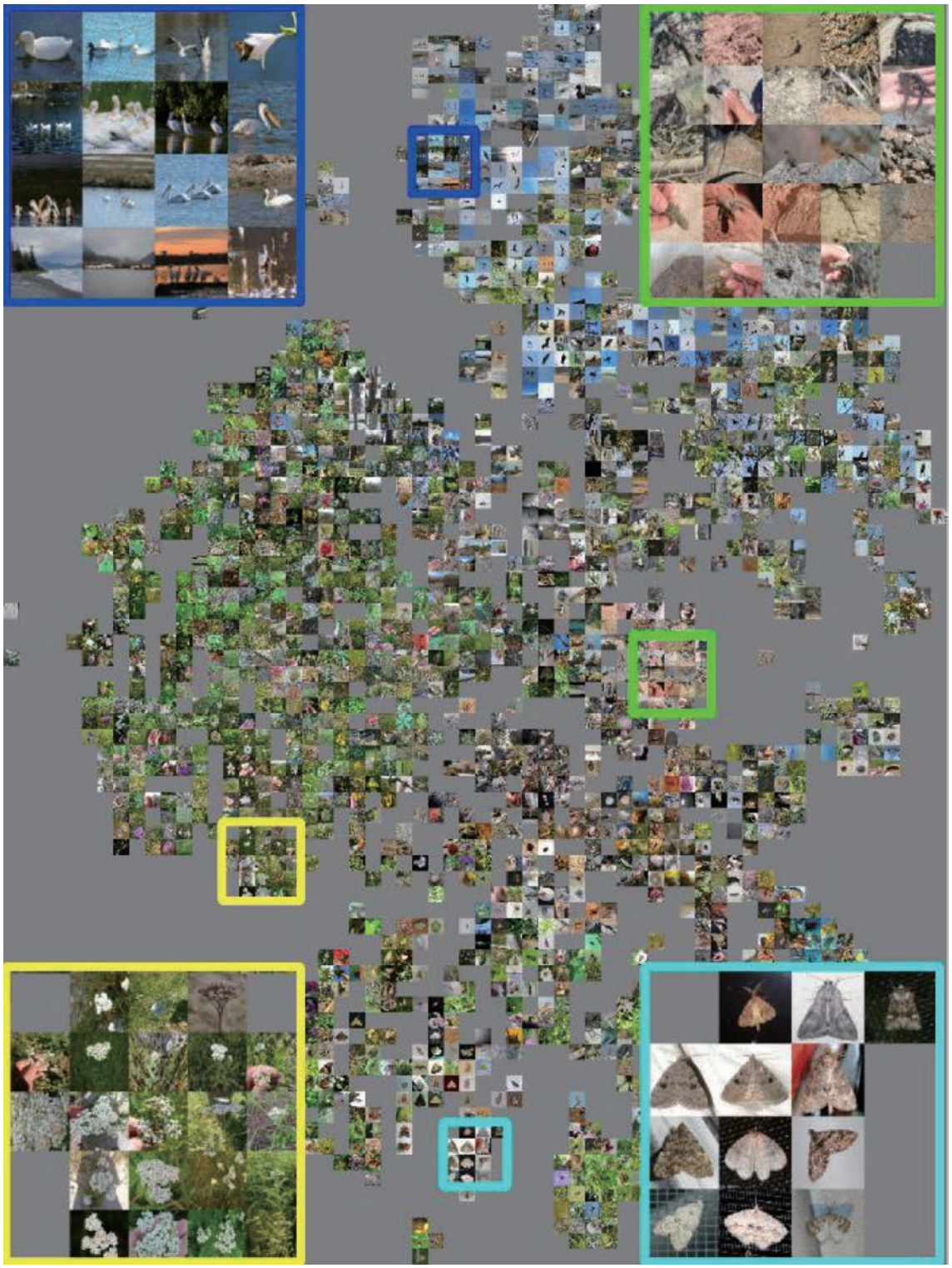
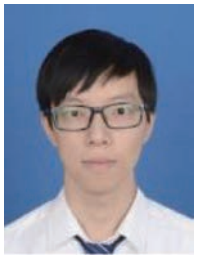


图 7 在 iNaturalist 数据集上嵌入可视化(在显示器上观看效果最佳)

D. 3. 目标检测结果示例

目标检测示例如图 8 所示。

图 8　MS COCO minival 测试集上的目标检测结果可视化(在显示器上查看效果最佳)



WEN Pei-Song, Ph. D., post doc fellow. His research interests include machine learning and computer vision, with special emphasis on learning to rank and self-supervised learning.

XU Qian-Qian, Ph. D., professor. Her research interests include statistical machine learning, with applications in multimedia and computer vision.

YANG Zhi-Yong, Ph. D., associate professor. His research interests lie in machine learning and learning theory, with special focus on AUC optimization, meta-learning/multi-task learning, and learning theory for recommender systems.

HUANG Qing-Ming, Ph. D., professor. His research areas include multimedia computing, image processing, computer vision and pattern recognition.

Background

This paper focuses on the Average Precision (AP) optimization, which is a foundation problem in machine learning. As an unbiased estimation of Area Under the Precision-Recall Curve (AUPRC), AP measures the trade-off between precision and recall, which is deemed as a more comprehensive metric, especially for highly skewed datasets. With the appealing property, AP is commonly adopted as a standard metric in various vision tasks like image retrieval, object detection, and long-tailed classification.

Although previous work has provided effective solutions for stochastic AP optimization, most of them focus on designing an approximation of the AP risk but ignore the generalization. Some early researches on information retrieval show that existing AP surrogate risks achieve an $O\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)$ generalization bound, where m and n are the number of queries and the length of the candidate list for a single query, respectively. This requires sufficient queries for a good generalization ability. However, for some real-world applications such as object detection and classification, only limited queries are available. In this case, how to analyze the generalization of an AP surrogate risk and further improve it is still an open problem.

To provide guidance for AP optimization, we present an early trial to study the generalization of AP loss within a single ranking list. The main challenge lies in the fact that AP risks cannot be decomposed into independent terms, making standard tools of generalization analysis infeasible. Besides, the relationship between commonly used surrogate losses and the original AP risk is unclear, making further analysis challenging.

To fill this gap, we start with the property of AP risks. First, we argue that the generalization is highly related to the stability of the AP surrogate objective. However, most existing surrogate objectives lack stability. This motivates us to propose a stable AP loss. Specifically, we start with a property of the AP loss: it can be viewed as a kind of ranking-weighted pairwise loss. Motivated by this fact, we derive an upper bound of the AP loss enjoying a ranking-weighted pairwise form. In this way, the AP loss is decomposed into stable components by estimating the weights, leading to provable generalization bounds in an order of $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$. To optimize the proposed surrogate loss efficiently, we first approximate the weights determined by the expectation of ranking. In this way, the surrogate objective is transformed into a two-level compositional function. Then we propose a stochastic optimization framework to jointly optimize the two levels. Theoretically, the proposed algorithm is proven to achieve an ϵ-stationary point with $O(1/\epsilon^4)$ complexity.

From the practical perspective, we conduct comprehensive empirical studies on seven benchmarks of image retrieval, object detection, and long-tailed classification, which show the broad application potential in machine learning and computer vision.