arXiv:2304.00395v1 [cs.LG] 1 Apr 2023

通过相似性结构理解对比学习机制：理论分析

Hiroki Waida[1] Yuichiro Wada[2,3]
waida.h.aa@m.titech.ac.jp wada.yuichiro@富士通.com L´eo And´eol[4,5,6,7] Takumi Nakagawa[1,3]
leo.andeol@math.univ-toulouse.fr nakagawa.t.as@m.titech.ac.jp
Yuhui Zhang[1] TakafumiKanamori[1,3]zhang.y.av@m.titech.ac.jp
kanamori@c.titech.ac.jp

[1]日本东京工业大学[2]日本富士通[3]日本理化学研究所人工智能项目[4]法国图卢兹数学研究所´法国[5]法国国家铁路公司[6]法国图卢兹大学´法国国家科学研究中心[7]

摘要

对比学习是一种高效的自监督表征学习方法。尽管近期研究在理论理解对比学习方面取得了进展，但对于如何刻画学习到的表示中的簇，相关研究仍然有限。本文旨在从理论角度阐明这种刻画。为此，我们考虑一种称为核对比学习（KCL）的基于核的对比学习框架，其中核函数在将我们的理论结果应用于其他框架时发挥着重要作用。我们从统计依赖的视角出发，提出一种刻画学习到的表示之相似性结构的方法。基于这一方法，我们研究了基于核的对比损失的理论性质。首先，我们证明，该方法能够刻画基于核的对比学习框架所学习到的表示结构。接着，我们给出下游任务分类误差的一个新上界，说明我们的理论与对比学习在实践中的成功相一致。我们还建立了KCL的泛化误差界。最后，我们通过一个替代界，给出KCL在下游分类任务上的泛化能力保证。

1 引言

近年来，许多关于自监督表征学习的研究都高度关注对比学习（Caron 等人，2020年；Chen 等人，2020a，b；Dwibedi 等人，2021年；HaoChen

等人，2021年；Li 等人，2021年）。通过对比学习，编码器函数利用数据背后的一些相似度信息，学会如何将无标签数据编码为优质表示。近期研究（Chen 等人，2020a，b；Dwibedi 等人，2021年）采用多种数据增强技术生成相似数据对。许多研究（Caron 等人，2020年；Chen 等人，2020a，b；Dwibedi 等人，2021年；HaoChen 等人，2021年）通过实证表明，对比学习能够生成有效的表示，并使其充分适用于分类和迁移学习等下游任务。

除了对比学习在实践中的发展之外，从理论上理解对比学习对于构建更高效的自监督学习算法也至关重要。在本文中，我们从理论角度探讨对比学习的以下基本问题：由对比学习预训练的编码器模型输出的特征向量簇具有怎样的特征？

近期，若干研究从多个理论视角探讨了这一问题，以研究对比学习在下游分类任务中的泛化保证（Arora 等，2019；Dufumier 等，2022；HaoChen 和 Ma，2023；HaoChen 等，2021；Huang 等，2023；Wang 等，2022a；Zhao 等，2023）。这些研究的主要方法之一是在数据中引入某种相似性度量。Arora 等人（2019）提出了条件独立性假设：给定从潜在类别分布中抽取的潜在类别 c，假设数据 x 和其正样本数据 x^+ 根据条件概率分布 $\mathcal{D}_c$独立采样。尽管潜在类别和条件独立性假设的概念常用于刻画 x 和x^+ 的语义相似性（Arora 等，2019；Ash 等，2022；Awasthi 等，2022；Bao 等，2022；Zou 和 Liu，2023），但若干研究（HaoChen 等，2021；Wang 等，2022a）指出，这一假设在实践中可能不成立。若干研究（HaoChen 等，2021；Wang 等，2022a）提出了关于数据间相似性的不同思路，以放宽这一假设。HaoChen 等人（2021）提出了总体增强图这一概念，在不对 x 和x^+作条件独立性假设的情况下，对谱对比学习（SCL）进行理论分析。一些研究还关注各种图结构（Dufumier 等，2022；HaoChen 和 Ma，2023；Wang 等，2022a）。尽管这些研究为理解对比学习提供了有趣的见解，但其分析的适用范围仅限于特定的目标函数。近期，若干研究（Huang 等，2023；Zhao 等，2023）考虑了这样一种设定：同一潜在类别中的原始数据彼此对齐良好，即增强距离足够小。尽管这些研究的理论保证适用于多种对比学习框架，但其对编码器函数类的假设较强，且这些保证在实践中能否成立仍有待阐明。因此，从统一视角研究上述问题仍在进行中，还需要更多视角来理解对比学习所学到的结构。

1.1 主要贡献

本文旨在通过引入一种基于数据间统计相似性的表述，从统一视角对上述问题进行理论研究。本文的主要贡献总结如下：

1. 由于我们旨在阐明对比学习的机制，因此需要考虑一个能够适用于其他方法的统一框架，而不是仅适用于 SimCLR（Chen 等人，2020a）和 SCL（HaoChen 等人，2021年）等特定框架。Li 等人（2021 年）指出，基于核的自监督学习目标与其他对比损失相关，例如 InfoNCE 损失（Chen 等人，2020a；van den Oord 等人，2018年）。因此，通过基于核的对比学习框架，可以从核的角度研究其他框架。受此启发，我们采用称为核对比学习（KCL）的框架作为实现这一目标的工具。KCL 的损失称为核对比损失，是一种形式简单且具有普适性的对比损失，其中两个特征向量之间的相似度通过再生核进行度量（Aronszajn，1950年；Berlinet and Thomas-Agnan，2004年；Steinwart andChristmann，2008年）（第 3 节）。我们的贡献之一是采用 KCL，从全新的统一理论视角研究对比学习的机制。

2. 我们提出了一种新的数据间相似性表述（第 4 节）。我们的相似性表述基于以下直觉：如果原始数据或增强数据 x 和x' 属于同一类别，那么由某个函数测得的相似度应高于某个阈值。基于这一点，我们提出了一种基于相似度函数（2）的表述（假设2）。

3. 我们对上述问题进行理论分析，以阐明其机制（第 5 节）。首先，我们证明 KCL 能够根据这一表述区分表示簇（第 5.1 节）。这一结果表明，我们的表述与对比学习的机制密切相关。接下来，我们为下游任务的分类误差建立了一个新的上界（第 5.2 节），这表明我们的表述并不与一系列工作所展示的对比学习在实践中的有效性相矛盾（Chen 等人，2020a，b；Dwibedi 等人，2021年；HaoChen 等人，2021年）。值得注意的是，与先前的工作（Huang 等人，2023年；Zhao 等人，2023年）相比，我们的上界在对编码器函数采用更现实的假设时仍然成立。我们还为 KCL 建立了泛化误差界（第 5.3 节）。最后，应用我们的理论结果，我们通过替代界给出了 KCL 泛化到下游分类任务的保证（第 5.4 节）。

1.2 相关工作

对比学习方法已从实证角度得到研究（Caron 等，2020；Chen 等，2020a、2021、2020b）。Chen 等人（2020a）提出了一种名为 SimCLR 的方法，该方法采用了 InfoNCE 的一个变体（Chen 等，2020a；van den Oord 等人，2018）。近期有多项研究从不同角度改进了对比方法（Caron 等，2020；Dwibedi 等人，2021；Robinson 等，2021a、b）。对比学习常用于多项基础任务，例如聚类（Van Gansbeke 等人，2020）和领域自适应（Singh，2021），也应用于视觉（Chen 等，2020a）、自然语言处理（Gao 等人，2021）和语音（Jiang 等人，2021）等领域。除对比方法外，也有多项研究（Chen 和 He，2021；Grill 等，2020）探讨了非对比方法。对比学习的理论理解也日益成为研究重点。例如，许多研究从不同设定出发，探讨了对比学习在下游分类任务上的泛化能力（Arora 等人，2019；Bao 等，2022；HaoChen 和 Ma，2023；HaoChen 等，2021；Huang 等，2023；Saunshi 等，2022；Tosh 等人，2021b；Wang 等，2022a；Zhao 等，2023）。

（Arora 等人，2019；Bao 等，2022；HaoChen 和 Ma，2023；HaoChen 等，2021；Huang 等，2023；Saunshi 等，2022；Tosh 等人，2021b；Wang 等，2022a；Zhao 等，2023）。多项研究从不同的理论和实证角度探讨对比学习，以阐明其机制，例如对比损失的几何性质（Huang 等，2023；Wang 和 Isola，2020）、数据间相似性的表述（Arora 等人，2019；Dufumier 等，2022；HaoChen 等，2021；Huang 等，2023；von K¨ugelgen 等人，2021；Wang 等，2022a；Zhao 等，2023）、归纳偏置（HaoChen 和 Ma，2023；Saunshi 等，2022）、迁移能力（HaoChen 等，2022；Shen 等，2022；Zhao 等，2023）、特征抑制（Chen 等，2021；Robinson 等，2021a）、负采样方法（Chuang 等，2020；Robinson 等，2021b）以及优化视角（Tian，2022；Wen 和 Li，2021）。

已有多项研究（Dufumier 等，2022；Johnson 等人，2023；Kiani 等，2022；Li 等，2021；Tsai 等，2022；Zhang 等，2022）探讨了对比学习与核理论之间的联系。Li 等人（2021）从核的角度研究了一些对比损失，例如 InfoNCE。Zhang 等，2022展示了核方法与 f互信息之间的关系，并将其理论应用于对比学习。Tsai 等，2022使用核作为相似度度量，解决了条件采样问题。Dufumier 等，2022考虑利用核函数理论将先验信息纳入对比学习。Kiani 等，2022通过优化问题的视角，将多种自监督学习算法与核方法联系起来。需要注意的是，与这些研究不同，我们的工作利用核函数，通过统计相似性探究对比学习的一种全新统一视角。

许多已有研究探讨了自监督表征学习目标的不同解释。例如，信息最大化原理（Poole 等，2019；Tschannen 等，2020）、谱聚类（HaoChen 等，2021）（关于谱聚类，参见 Ng 等人（2002））以及希尔伯特-施密特独立性准则（HSIC）（Li 等，2021）（关于 HSIC，参见 Gretton（2005））。然而，从统一视角研究对比学习值得深入探讨，以阐明其机制，因为近期关于自监督表征学习的研究从不同立场对其进行了研究（Dubois 等，2022；Huang 等，2023；Johnson 等人，2023；Kiani 等，2022；Tian，2022）。

2 预备知识

2.1 问题设定

我们给出对比学习的标准设定。我们的设定与 HaoChen 等人（2021）的设定非常接近，不过我们还引入了额外的技术性必要设定，以保证数学上的严谨性。设 $\mathbb{X} \subset \mathbb{R}^p$ 是由原始数据构成的拓扑空间，$P_{\mathbb{X}}$ 是 $\mathbb{X}$ 上的 Borel 概率测度。一系列对比学习研究（Chen 等，2020a,b；Dwibedi 等人，2021；HaoChen 等人，2021）使用数据增强技术来获得相似的增强数据点。因此，我们定义一组 $\mathcal{T}$，其中的映射将点 $x \in \mathbb{X}$ 变换为 $\mathbb{R}^p$，并假设 $\mathcal{T}$ 包含 $\mathbb{R}^p$ 上的恒等映射。接下来，定义 $\mathbb{X} = \bigcup_{t\in\mathcal{T}}\{t(x): \ x \in \mathbb{X}\}$。$\mathcal{T}$ 中的每个元素 t 都可视为一个映射，它将原始数据点 $x \in \mathbb{X}$ 返回为增强数据 $x = t(x)$。注意，由于恒等映射属于 $\mathcal{T}$，因此 $\mathbb{X}$ 是 $\mathbb{X}$ 的子集。我们为 $\mathbb{X}$ 配备某种拓扑。设 $\nu_{\mathbb{X}}$ 是 $\mathbb{X} \subset \mathbb{R}^p$ 上的一个 σ- 有限非负 Borel

测度。遵循 HaoChen 等人（2021）的思路，我们将 $a(x|x)$ 记为给定 $x \sim P_{\mathbb{X}}$ 时 x 的条件概率密度函数，并将权重函数定义为 $w : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$，即 $w(x, x') = \mathbb{E}_{x \sim P_{\mathbb{X}}}[a(x|x)a(x'|x)]$。根据定义，$w$ 是 $\mathbb{X} \times \mathbb{X}$ 上的联合概率密度函数。我们将权重函数的边缘分布 $w(\cdot)$ 定义为 $w(x) = \int w(x, x') d\nu_{\mathbb{X}}(x')$。边缘分布 $w(\cdot)$ 也是 $\mathbb{X}$ 上的概率密度函数，相应的概率测度记为 $dP_{\mathbb{X}}(x) = w(x)d\nu_{\mathbb{X}}(x)$。分别用 $\mathbb{E}_{x, x^+}[\cdot]$、$\mathbb{E}_{x, x^-}[\cdot]$ 表示关于概率测度 $w(x, x')d\nu_{\mathbb{X}}^{\otimes 2}(x, x')$ $w(x)w(x')d\nu_{\mathbb{X}}^{\otimes 2}(x, x')$ 在 $\mathbb{X} \times \mathbb{X}$ 上的期望，其中 $\nu_{\mathbb{X}}^{\otimes 2} := \nu_{\mathbb{X}} \otimes \nu_{\mathbb{X}}$ 是 $\mathbb{X} \times \mathbb{X}$ 上的乘积测度。为严谨地表述我们的对比学习框架，我们假设边缘分布 w 在 $\mathbb{X}$ 上为正。实际上，满足 $w(x) = 0$ 的点 $x \in \mathbb{X}$ 不属于 $a(\cdot|x)$ 的支撑集，且这一结论对 $P_{\mathbb{X}}$-几乎处处的 $x \in \mathbb{X}$ 均成立，这意味着这样的点 x 仅作为增强数据出现。

设 $f_0 : \mathbb{X} \to \mathbb{R}^d$ 是一个将增强数据映射到特征空间的编码器模型，并设 $\mathcal{F}_0$ 是由此类编码器构成的函数类。实践中，f_0 由骨干网络架构（例如，残差网络（He 等人，2016）；参见 Chen 等人（2020a））定义，后接称为投影头的额外多层感知机（Chen 等人，2020a）。我们假设 $\mathcal{F}_0$ 一致有界，即存在一个通用常数 $c \in \mathbb{R}$，使得 $\sup_{f_0 \in \mathcal{F}_0} \sup_{x \in \mathbb{X}} \|f_0(x)\|_2 \leq c$。例如，在有界域上由无偏置全连接神经网络构成的函数空间中，若每个神经网络的各层都具有连续激活函数，则该函数空间满足此条件。由于在许多实证研究（Chen 等人，2020a；Dwibedi 等人，2021年）和若干理论研究（Wang 和 Isola，2020；Wang 等，2022a）中，编码器模型输出的特征向量都使用欧几里得范数进行归一化，因此我们考虑归一化函数的函数空间 $\mathcal{F} = \{f \mid \exists f_0 \in \mathcal{F}_0,\ f(x) = f_0(x)/\|f_0(x)\|_2$，其中 $\forall x \in \mathbb{X}\}$。这里，为保证每个 $f \in \mathcal{F}$ 都有定义，假设 $\mathfrak{m}(\mathcal{F}_0) := \inf_{f \in \mathcal{F}} \inf_{x \in \mathbb{X}} \|f_0(x)\|_2 > 0$ 成立。

最后，我们介绍本文中通用的若干记号。设 $\mathbb{M} \subset \mathbb{X}$ 是一个可测集，则记作

$$\mathbb{E}[f(x)|\mathbb{M}] := \int_{\mathbb{X}} f(x) P_{\mathbb{X}}(dx|\mathbb{M}) = P_{\mathbb{X}}(\mathbb{M})^{-1} \int_{\mathbb{M}} f(x) w(x) d\nu_{\mathbb{X}}(x).$$

我们还使用记号 $\mathbb{E}[f(x); \mathbb{M}] := \int_{\mathbb{M}} f(x)w(x)d\nu_{\mathbb{X}}(x)$。用 $\mathbb{1}_{\mathbb{M}}(\cdot)$表示集合 $\mathbb{M}$的示性函数。我们还用 $[n] := \{1, \cdots, n\}$ 表示 $n \in \mathbb{N}$。

2.2 再生核

我们介绍再生核的一些记号（Aronszajn，1950年；Berlinet and Thomas-Agnan，2004年；Steinwart and Christmann，2008年）。设 k：$\mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ 是一个实值、连续、对称且正定的核，其中 $\mathbb{S}^{d-1}$ 表示以原点 $\mathbf{0} \in \mathbb{R}^d$为中心的单位超球面，正定性意味着对于任意 $\{z_i\}_{i=1}^n \subset \mathbb{S}^{d-1}$ 和 $\{c_i\}_{i=1}^n \subset \mathbb{R}$，$\sum_{i,j=1}^n c_i c_j k(z_i, z_j) \geq 0$ 均成立（Berlinet and Thomas-Agnan，2004年）。设 $\mathcal{H}_k$ 是以核 k 为核函数的再生核希尔伯特空间（RKHS）（Aronszajn，1950年），它满足 $\phi(z) = \langle \phi, k(\cdot, z) \rangle_{\mathcal{H}_k}$，对所有 $\phi \in \mathcal{H}_k$ 和 $z \in \mathbb{S}^{d-1}$均成立。对于 $z \in \mathbb{S}^{d-1}$，记 $h(z) = k(\cdot, z)$；这种映射通常称为特征映射（Steinwart and Christmann，2008年）。这里，我们施加以下条件。

假设 1。对于核函数 k，存在某个 ρ-利普希茨函数 ψ：$[-1, 1] \to \mathbb{R}$，使得对于任意 z、$z' \in \mathbb{S}^{d-1}$、$k(z, z') = \psi(z^\top z')$ 均成立。

机器学习中的几种常用核函数，例如线性核、二次核和高斯核，都满足假设1。注意，假设 1 中的利普希茨条件有助于推导核对比损失的泛化误差界（见第5.3节），并且在针对特定核函数进行分析时，有时可以去除该条件。我们使用这一假设来给出一般性结果。本文还使用以下概念：

命题 1. 设 $\mathbb{M} \subset \mathbb{X}$ 为一个可测集，且 $f \in \mathcal{F}$。定义 $\mu_{\mathbb{M}}(f) := \mathbb{E}_{P_{\mathbb{X}}}[h(f(x))|\mathbb{M}]$。则，$\mu_{\mathbb{M}}(f) \in \mathcal{H}_k$。

量 $\mu_{\mathbb{M}}(f$在可测子集 $\mathbb{M}$ 下可视为核均值嵌入的一种变体（Muandet 等人，2017年）。命题 1 的证明是对 Muandet 等人 (2017 年的略微修改；见附录 A。

3 核对比学习

本节介绍一个对比学习框架，用于分析对比学习的机制。在表征学习中，InfoNCE 损失（Chen 等人，2020a；van den Oord 等人，2018年）广泛应用于视觉（Chen 等人，2020a、2021年；Dwibedi 等人，2021年）等应用领域。遵循以往的研究（Bao 等人，2022年；Chen 等人，2020a；van den Oord 等人，2018年），我们将 InfoNCE 损失定义为：

$$L_{\mathrm{NCE}}(f;\tau) = -\mathbb{E}_{\substack{x,x^+ \\ \{x_i^-\}\sim P_{\mathbb{X}}}} \left[\log \frac{e^{f(x)^\top f(x^+)/\tau}}{e^{f(x)^\top f(x^+)/\tau} + \sum_{i=1}^{M} e^{f(x)^\top f(x_i^-)/\tau}} \right],$$

其中 $\{x_i^-\}$ 是独立同分布随机变量，$\tau > 0$，且 M 是负样本的数量。Wang 和 Isola (2020年) 给出了 InfoNCE 损失的渐近形式：

$$L_{\infty\text{-NCE}}(f;\tau) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \mathbb{E}_x\left[\log \mathbb{E}_{x'}\left[e^{\frac{f(x)^\top f(x')}{\tau}}\right]\right].$$

根据 Wang 和 Isola（2020年）的理论分析，他们指出，第一项表示对齐，即特征向量对 (x, x^+) 之间的平均接近程度；第二项则表示均匀性，即负样本 x, x' 的特征向量彼此相距多远。此外，Chen 等人（2021年）报告了广义对比损失的有效性，这类损失还包含额外的权重超参数。同时，由于我们旨在研究对比学习的机制，因此需要一种与其他框架相关的简单且通用的对比损失形式。在此，Li 等人（2021年）通过表明某些 HSIC 准则与目标函数$\mathbb{E}_{x,x^+}[k(f(x)\ f(x^+))] - \mathbb{E}_{x,x^-}[k(f(x)\ f(x^-))]$成正比，发现了自监督学习与核之间的联系（更多细节见附录F.2）。受这一联系启发，我们考虑一种对比学习目标，其中核函数用于衡量增强数据点的特征向量之间的相似度。更具体地说，对于第2.2节中引入的核函数k，我们将核对比损失定义为：

$$L_{\mathrm{KCL}}(f;\lambda) = -\mathbb{E}_{x,x^+}\left[k(f(x), f(x^+))\right] + \lambda \mathbb{E}_{x,x^-}\left[k(f(x), f(x^-))\right],$$

其中，权重超参数 λ的灵感来自 Chen 等人（2021年）。在此，核对比损失 L_{KCL} 在对比学习的预训练阶段被最小化。在本文中，

采用核对比损失的对比学习框架称为核对比学习（KCL）。

接下来，我们展示它与其他对比学习目标之间的联系。首先，对于 InfoNCE 损失，我们考虑通过选择 $L_{\mathrm{LinKCL}}(f;\ \lambda)$ 定义的线性核对比损失 $k(z,\ z') = z^\top z'$。请注意，L_{LinKCL} 以及其经验损失也在若干研究中有所讨论（Huang 等人，2023年；Wang 和 Liu，2021年）。对于 $L_{\mathrm{LinKCL}}(f;\ 1)$，我们有：

$$\tau^{-1} L_{\mathrm{LinKCL}}(f;1) \le L_{\mathrm{NCE}}(f;\tau) + \log M^{-1}. \tag{1}$$

在附录F.3中，我们给出了针对广义损失（Chen 等人，2021年）的（1）的广义不等式。请注意，当用渐近损失（ L_{NCE} 替代时，相似的关系仍然成立（Wang 和 Isola，2020年）或解耦对比学习损失（ Yeh 等人，2022年）；参见附录 F.3。因此，可以通过 $L_{\mathrm{LinKCL}}(f;\ \lambda)$ 分析 InfoNCE 损失及其变体。

核对比损失也与其他对比学习目标有关。例如，采用二次核的二次核对比损失 k（$z,\ z'$）$=$ （$z^\top z'$）^{2} 在加上一个常数后成为谱对比损失（HaoChen 等人，2021年）的一个下界（见附录F.3）。因此，对核对比损失的理论分析可以应用于其他对比学习目标。

请注意，我们通过实验证明，采用高斯核和二次核的 KCL 框架虽然简单，但确实有效；实验设置见附录H.1，补充材料中的实验结果见附录H.2 和H.3 。实验结果也促使我们使用 KCL 作为研究对比学习的理论工具。

4 基于统计相似度的表述

4.1 关键要素：相似度函数

为了研究对比学习的机制，我们引入增强数据点之间相似度的概念，这是我们分析中的关键组成部分。定义如下：

$$\mathrm{sim}(x,x';\lambda) := \frac{w(x,x')}{w(x)w(x')} - \lambda, \tag{2}$$

其中，$\lambda \ge 0$ 是 L_{KCL}的权重参数，$w(x,x')$和 $w(x)$已在第 2.1 节中介绍。注意，式 (2) 是良定义的，因为对每个 $x \in \mathbb{X}$，$w(x) > 0$ 均成立。量 $\mathrm{sim}(x,x';\lambda)$表示 x 和 x' 之间统计依赖的强弱。密度比 $w(x,x')/(w(x)w(x'))$可视为 Tsai 等人（2020年）所介绍的一种逐点依赖。超参数 λ 通过 x和 x' 之间的（不）依赖关系，控制两份增强数据之间的相关程度。例如，在固定 $\lambda = 1$的情况下，如果$w(x,x') > w(x)w(x')$，那么 $\mathrm{sim}(x,x';1)$ 为正，即 x 和 x' 相关。

这里需要指出，表征学习领域的若干理论研究（Johnson 等人，2023年；Tosh 等人，2021b；Wang 等人，2022b）使用密度比来研究若干对比学习目标的最优表示。 Tosh 等人 (2021b）关注逻辑损失的极小化器可表示为密度比这一事实，并利用这一点研究 里程碑式的

嵌入 (Tosh 等人，2021a）。 Johnson 等人(2023年）将密度比与若干对比学习目标的极小化器联系起来，并考察与这些极小化器相关的表示的质量。此外，Wang 等人(2022b）研究了谱对比损失的极小化器（HaoChen 等人，2021年）。与此同时，我们强调，在（2）中使用密度比的目的并非研究 KCL 的最优表示，而是给出一种基于增强数据之间统计相似性的表述。

注 1. 我们可以证明，核对比损失可视为总体归一化割问题（*Terada 和 Yamamoto，2019年*）的一种松弛，其中积分核由 (2) 定义。因此，(2) 定义了 *KCL* 所采用的相似性结构。详细论证以及与相关工作的比较（*HaoChen 等人，2021年*；*Tian，2022年*）见附录 *G*。

4.2 公式化与示例

我们根据问题设定引入以下公式化表述。

假设 2. 存在某个 $\delta \in \mathbb{R}$、簇数 $K \in \mathbb{N}$、可测子集$\mathbb{M}_1$、$\cdots$、$\mathbb{M}_K \subset \mathbb{X}$以及确定性标注函数 $y\colon \mathbb{X} \to [K]$，使得以下条件成立：

(A) $\bigcup_{i=1}^{K} \mathbb{M}_i = \mathbb{X}$ 成立。**(B)** 对于每个 $i \in [K]$，任意点 x、$x' \in \mathbb{M}_i$ 均满足 $\mathrm{sim}(x, x'; \lambda) \geq \delta$。**(C)** 对于每个 $x \in \mathbb{X}$ 以及指标集合 $J_x = \{j \in [K] \mid x \in \mathbb{M}_j\}$、$y(x) \in J_x$ 成立。此外，每个集合 $\{x \in \mathbb{X} \mid y(x) = i\}$ 都是可测的。

假设 2 并不要求 $\mathbb{M}_1$、$\cdots$、$\mathbb{M}_K$ 两两不相交；这是一个符合实际的设定，因为 Wang 等人(2022a) 表明，增强数据的簇可能会因数据增强强度不同而存在簇间连接。假设 **(A)** 和 **(B)** 保证假设 2 中的每个子集 $\mathbb{M}_i$ 都由相似度较高的增强数据构成。条件 **(C)**使我们能够在分析中纳入标签信息。值得注意的是，若干关于对比学习的研究（HaoChen 和 Ma，2023年；HaoChen 等人，2021年；Saunshi 等人，2022年）也采用确定性标注函数。

这些条件有助于分析对比学习的理论。为了增进直观理解，我们给出一个满足假设2的简单例子。

例 *1*（证明见附录 *F.1*）。设 $\mathbb{X}$ 由半径相同且两两不相交的开球 $\mathbb{B}_1, \cdots, \mathbb{B}_K$ 构成，这些开球位于 $\mathbb{R}^p$中，并且 $\mathbb{X} = \mathbb{X}$。令 $a(x|x) = vol(\mathbb{B}_1)^{-1} \sum_{i=1}^{K} \mathbb{1}_{\mathbb{B}_i \times \mathbb{B}_i}(x, x)$，并令$p_{\mathbb{X}}(x) = (Kvol(\mathbb{B}_1))^{-1}$ 为 $\overline{P}_{\mathbb{X}}$ 的概率密度函数。则， $w(x) > 0$ 且 $sim(x, x'; \lambda) = K\mathbb{1}_{\bigcup_{i\in[K]} \mathbb{B}_i \times \mathbb{B}_i}(x, x') - \lambda$ 成立。因此，例如，令 $\lambda = 1$ 和 $\delta = K - 1$，并取 $\mathbb{M}_i := \mathbb{B}_i$。此外，定义 $y\colon \mathbb{X} \to [K]$，使得当 $x \in \mathbb{B}_i$ 对某个 $i \in [K]$成立时，$y(x) = i$。那么，在此设定下，假设 *2* 得以满足。

在此，已有多项关于对比学习的研究探讨了相似度的理论表述（Arora 等人，2019年；Dufumier 等人，2022年；HaoChen 和 Ma，2023年；HaoChen

等人，2021年；Huang 等人，2023年；Parulekar 等人，2023年；von K¨ugelgen 等人，2021年；Wang 等人，2022a；Zhao 等人，2023年）。这些研究引入的基本概念包括：无标签数据的潜在类别与条件独立性假设（Arora 等，2019）、增强数据的图结构（Dufumier 等，2022；HaoChen 和 Ma，2023；HaoChen 等，2021；Wang 等，2022a）、将无标签数据分解为内容变量（即对数据增强保持不变）和风格变量（即会随数据增强而改变）（Parulekar 等人，2023；von K¨ugelgen 等人，2021），以及基于增强距离和类别子集内集中性的几何结构（Huang 等，2023；Zhao 等，2023）。与此同时，我们在假设 2 中采用了相似度函数（2），这与先前的研究有所不同。值得注意的是，假设 2 与 HaoChen 和 Ma（2023）中的假设 3 存在一定联系；参见附录 F.4。如第 5 节所示，我们的表述为理解对比学习提供了更深入的见解。

5 理论结果

5.1 基于统计相似性的 KCLas表示学习

首先，我们将核对比损失与基于相似度 $\mathrm{sim}(x, x'; \lambda)$ 的公式联系起来。下面的定理表明，KCL 对表征学习的方式具有两种影响，其中簇 $\mathbb{M}_1, \cdots, \mathbb{M}_K$ 参与解释这一数学关系。

定理1。 假设 *1* 和 *2* 成立。取 $\delta \in \mathbb{R}$, $K \in \mathbb{N}$、 $\mathbb{M}_1, \cdots, \mathbb{M}_K$，使得假设 **2** 中的条件 **(A)** 和 *(B)* 得到满足。则对于每个 $f \in \mathcal{F}$:

$$\frac{\delta}{2} \cdot \mathfrak{a}(f) + \lambda \cdot \mathfrak{c}(f) \leq L_{\mathrm{KCL}}(f; \lambda) + R(\lambda). \tag{3}$$

其中， $R(\lambda)$ 是 λ, $\mu_i(f) = \mu_{\mathbb{M}_i}(f)$的函数，并且

$$\mathfrak{a}(f) := \sum_{i=1}^{K} \mathbb{E}_{x,x^-} \left[\|h(f(x)) - h(f(x^-))\|^2_{\mathcal{H}_k}; \mathbb{M}_i \times \mathbb{M}_i \right],$$
$$\mathfrak{c}(f) := \sum_{i \neq j} P_{\mathbb{X}}(\mathbb{M}_i) P_{\mathbb{X}}(\mathbb{M}_j) \langle \mu_i(f), \mu_j(f) \rangle_{\mathcal{H}_k}.$$

定理1的证明见附录 B.2。需要注意，证明的关键在于如下运用假设 2：对于任意 $x, x' \in \mathbb{M}_i$, 不等式 $\mathrm{sim}(x, x'; \lambda) \geq \delta$蕴含关系 $w(x, x') \geq (\lambda + \delta) w(x) w(x')$。下面简要解释定理1中的各个符号。值 $\mathfrak{a}(f)$量化了由 $\mathbb{M}_i$ 中增强数据的表示构成的每个簇内的集中程度。量 $\mathfrak{c}(f)$衡量子集 $h(f(\mathbb{M}_1)) \cdots, h(f(\mathbb{M}_K))$之间的距离，因为 $\langle \mu_i(f) \mu_j(f) \rangle_{\mathcal{H}_k} = \int_{\mathbb{M}_i} \int_{\mathbb{M}_j} k(f(x) \ f(x')) P_{\mathbb{X}}(dx|\mathbb{M}_i) P_{\mathbb{X}}(dx'|\mathbb{M}_j)$ 成立（见附录 B.1 中的引理 2）。这些量表明，KCL 的表征学习能够在再生核希尔伯特空间中区分子集 $\mathbb{M}_1, \cdots, \mathbb{M}_K$（示意见图1）。函数 $R(\lambda)$包含一项，用以表示增强数据空间 $\mathbb{X}$ 中预训练任务的难度：如果两个不同子集 $\mathbb{M}_i$ 和 $\mathbb{M}_j$ 之间的重叠扩大，那么 $R(\lambda)$就会增大（精确定义见附录 B.2）。

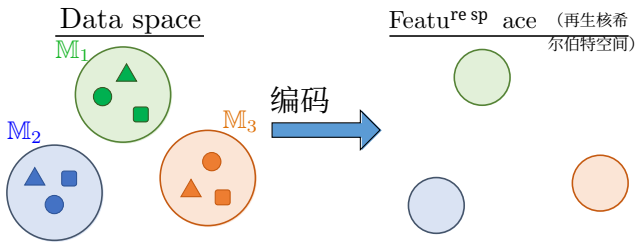


图1：定理1的示意图。簇 $\mathbb{M}_1$、$\cdots$、$\mathbb{M}_K$ 在数据空间中被映射到再生核希尔伯特空间（RKHS）中，其中每个簇在再生核希尔伯特空间（RKHS）中收缩或扩张（通过 $\mathfrak{a}(f)$），同时保持与其他簇的距离（通过 $\mathfrak{c}(f)$）。

定理1的一个关键点是，δ 和 λ 可以决定学习到的表征在再生核希尔伯特空间中的分布。如果 $\delta > 0$，那么每个 $\mathbb{M}_i$ 中的增强数据表示倾向于对齐，同时调节 $\mathfrak{a}(f)$ 与 $\mathfrak{c}(f)$ 之间的权衡。对于 $\delta \leq 0$，不仅均值 $\mu_1(f) \cdots, \mu_K(f)$，表示本身也倾向于分散。我们指出，由于假设 2 中的条件 (B)，δ 取决于固定权重 λ。直观而言，较大的 λ 会使 δ 变小，反之亦然。

这里需要指出，在若干假设下，等式（3）成立，如下所示：

推论*1.* 假设假设 *1* 和 *2* 成立。取 $\delta \in \mathbb{R}$、$K \in \mathbb{N}$，以及 $\mathbb{M}_1$、$\cdots$、$\mathbb{M}_K$，使得满足假设 *2* 中的条件 *(A)* 和 *(B)*。假设 $\mathbb{M}_1$、$\cdots$、$\mathbb{M}_K$ 两两不相交，并且对于每一对 $(i, j) \in [K] \times [K]$，满足 $i \neq j$，每个 $(x, x') \in \mathbb{M}_i \times \mathbb{M}_j$ 均满足 $w(x, x') = 0$。假设对于每个 $i \in [K]$，对于任意 x、$x' \in \mathbb{M}_i$，均有 $sim(x, x'; \lambda) = \delta$。那么，对于每个 $f \in \mathcal{F}$，等式 *(3)* 成立，即

$$L_{\mathrm{KCL}}(f; \lambda) = \frac{\delta}{2} \cdot \mathfrak{a}(f) + \lambda \cdot \mathfrak{c}(f) - R(\lambda).$$

推论1的证明见附录B.3。上述推论意味着，在这些假设下，最小化核对比损失等价于最小化目标函数 $(\delta/2) \cdot \mathfrak{a}(f) + \lambda \cdot \mathfrak{c}(f)$。注意，示例1满足陈述中列出的所有假设。总之，定理1和推论1表明，KCL框架下的对比学习可被刻画为基于增强数据空间相似性结构的表征学习 $\mathbb{X}$。

5.1.1 与相关工作的比较

量 $\mathfrak{a}(f)$ 与称为对齐的性质（Wang 和 Isola，2020）密切相关，因为相似数据的表示是在再生核希尔伯特空间中学习为彼此接近的。此外，量 $\mathfrak{c}(f)$ 与发散性（Huang 等人，2023）有一定联系，因为它衡量均值 $\mu_i(f)$ 和 $\mu_j(f)$ 之间的距离。尽管 Huang 等人（2023）和 Wang 和 Isola（2020）已指出这些性质与对比学习之间的关系，但我们强调，我们的结果为学习到的簇提供了一种新的刻画。此外，该定理还表明，$\mathfrak{a}(f)$ 与 $\mathfrak{c}(f)$ 之间的权衡由阈值 δ 和超参数 λ 决定。

因此，定理 1为深入理解对比学习的机制提供了更深刻的见解。

5.2 分类误差的新上界

接下来，根据我们对相似度的表述，我们说明最小化核对比损失如何保证下游分类任务取得良好表现。为此，我们证明，对比学习在定理1 中展现的性质能够在再生核希尔伯特空间中产生线性良好可分表示。首先，我们如下量化线性可分性：沿用 HaoChen 等人 (2021年); Saunshi 等人 (2022年), 在假设 2下，对于模型 $f \in \mathcal{F}$、线性权重 W： $\mathcal{H}_k \to \mathbb{R}^K$和偏置 $\beta \in \mathbb{R}^K$，我们将下游分类误差定义为：

$$L_{\mathrm{Err}}(f, W, \beta; y) := P_{\mathbb{X}}\left(g_{f,W,\beta}(x) \neq y(x)\right),$$

其中 $g_{f,W,\beta}(x) :=$ 最大值自变量$_{i\in[K]}\{\langle W_i, h(f(x))\rangle_{\mathcal{H}_k} + \beta_i\}$ 适用于 $W_i \in \mathcal{H}_k$和 $\beta_i \in \mathbb{R}$。注意，我们与 HaoChen 等人（2021年）一样，令最大值自变量和最小值自变量在出现并列时任意选择。还要注意，在我们的定义中，增强数据 $x \in \mathbb{X}$ 被编码为 $f(x) \in \mathbb{S}^{d-1}$后， $f(x)$ 会进一步映射为$h(f(x))$，进入再生核希尔伯特空间，然后使用 W 和 β进行线性分类。

为了推导 KCL 的泛化保证，我们将再生核希尔伯特空间中的 1-最近邻（NN）分类器作为 $\mathcal{H}_k$ 一种技术工具；它是以下文献所用 1-NN 分类器的推广：Huang 等人 (2023年); Robinson 等人(2021b).

定义 1 （再生核希尔伯特空间中的 1-NN 分类器 $\mathcal{H}_k$）。设 $\mathbb{C}_1$、 $\cdots$、 $\mathbb{C}_K$ 是 $\mathbb{X}$ 的子集。对于模型 f: $\mathbb{X} \to \mathbb{R}^d$, 与再生核希尔伯特空间 $\mathcal{H}_k$ 相关联的 *1-NN* 分类器 $g_{1\text{-NN}}$： $\mathbb{X} \to [K]$ 定义为

$$g_{1\text{-NN}}(x) := \arg\min_{i\in[K]} \|h(f(x)) - \mu_{\mathbb{C}_i}(f)\|_{\mathcal{H}_k}.$$

Huang 等人（2023年）表明，他们所考虑的 1-NN 分类器可视为均值分类器（Arora 等人，2019年; Wang 等人，2022a）（参见Huang 等人 （2023年）中的附录 E）。这一事实也适用于我们的设定：事实上，在假设1下， $g_{1\text{-NN}}$ 等于

$$g_{f,W_\mu,\beta_\mu}(x) = \arg\max_{i\in[K]} \left\{\langle W_{\mu,i}, h(f(x))\rangle_{\mathcal{H}_k} - \beta_{\mu,i}\right\},$$

其中 W_μ: $\mathcal{H}_k \to \mathbb{R}^K$定义为 $W_\mu(\phi)_i := \langle W_{\mu,i}, \phi\rangle_{\mathcal{H}_k} = \langle \mu_{\mathbb{C}_i}(f)\ \phi\rangle_{\mathcal{H}_k}$对每个坐标$i \in [K]$, 且 $\beta_{\mu,\ i}$: $= (\|\mu_{\mathbb{C}_i}(f)\|^2_{\mathcal{H}_k} + \psi(1))/2$ 对于 $i \in [K]$。

在给出结果之前，我们需要以下概念：

定义 2 (有意义的编码器)。 如果编码器 $f \in \mathcal{F}$满足$\min_{i\neq j} \|\mu_i(f) - \mu_j(f)\|^2_{\mathcal{H}_k} > 0$ ， 则称其为有意义的编码器。

请注意，有意义的编码器 $f \in \mathcal{F}$ 可以避免特征向量的完全坍塌（Hua 等人，2021年; Jing 等人，2022年），许多关于自监督表征学习的研究（Chen 等人，2020a; Chen 和 He，2021年; Grill 等人，2020年; HaoChen 等人，2021年; Li 等人，2021年）提出了各种架构和算法来防止这种情况。现给出理论保证：

定理 2。 假设假设 *1*和 *2*成立。取 $\delta \in \mathbb{R}$， $K \in \mathbb{N}$， $\mathbb{M}_1$，$\cdots$，$\mathbb{M}_K$，以及 y，使得假设 **(A)**、**(B)**和 **(C)** 中的条件 *2* 均满足。于是，对于每个有意义的编码器 $f \in \mathcal{F}$，有

$$L_{\mathrm{Err}}(f, W_\mu, \beta_\mu; y) \leq \frac{8(K-1)}{\Delta_{\min}(f) \cdot \min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)} \mathfrak{a}(f)$$

其中 $\Delta_{\min}(f) = \min_{i\neq j} \|\mu_i(f) - \mu_j(f)\|^2_{\mathcal{H}_k}$。

定理 2 的证明见附录 C。如果任意两个属于 $\mathbb{M}_i$ 的点 x, x' 的表示对于每个 $i \in [K]$ 都更接近，并且不同子集 $\mathbb{M}_i$ 和 $\mathbb{M}_j$ 的最近中心 $\mu_i(f)$ 与 $\mu_j(f)$ 彼此距离更远，那么定理 2 中的上界就会变小。由于 $\|\mu_i(f) - \mu_j(f)\|^2_{\mathcal{H}_k} = -2\langle \mu_i(f), \mu_j(f)\rangle_{\mathcal{H}_k} + \|\mu_i(f)\|^2_{\mathcal{H}_k} + \|\mu_j(f)\|^2_{\mathcal{H}_k}$，定理1和定理 2 表明，在优化核对比损失的过程中，量 $\mathfrak{a}(f)$ 和 $\mathfrak{c}(f)$ 有助于使学习到的表征在线性意义上具有良好的可分性。因此，我们的理论与一系列研究所揭示的对比学习的实证成功相一致（Chen 等，2020a,b；Dwibedi 等人，2021；HaoChen 等，2021）。

5.2.1 与相关工作的比较

我们讨论定理 2。1）一些研究（Huang 等，2023；Robinson 等，2021b）也表明，分类损失或误差有一个上界，该上界与每个簇内特征向量的对齐程度有关。然而，他们的结果没有探讨以下猜想：由特征向量构成的各个簇的中心之间的距离是否会影响线性可分性？定理 2 表明，答案是肯定的，这可通过量 $\Delta_{\min}(f)$ 得出。注意，Zhao 等人（2023）的定理 3.2 也蕴含了类似的答案，但他们的结果针对平方损失，且对编码器函数提出了若干强假设。与此同时，我们关于分类误差的定理 2 要求编码器函数具有有意义性（定义 2），这比 Zhao 等人（2023）所要求的更具实践性。2）此外，我们的结果在问题设定上不同于以往研究（Huang 等，2023；Robinson 等，2021b；Zhao 等，2023）。具体而言，Robinson 等（2021b）遵循 Arora 等人（2019）的设定，而 Huang 等（2023）和 Zhao 等（2023）则通过对给定的潜在类别子集施加（σ, δ）-增强性质来构建其设定。相比之下，我们的公式主要基于统计相似性（2）。此外，我们指出，我们的定理 2 可以推广到 $K \in \mathbb{N}, \mathbb{M}_1, \cdots, \mathbb{M}_K \subset \mathbb{X}$ 和 y 取值满足假设 2 中条件 (A) 和 (C) 的情形（见附录 C 中的定理 5），这表明我们的结果适用于对比学习的其他问题设定。受篇幅限制，我们在附录 F.6 和 F.7 中给出更详细的解释。

5.3 KCL 的泛化误差界

由于实践中我们最小化经验核对比损失，因此推导出 KCL 的泛化误差界。经验损失定义如下：记 $dP_+(x, x') = w(x, x')d\nu^{\otimes 2}_{\mathbb{X}}(x, x')$。令 $(X_1, X'_1) \cdots, (X_n, X'_n)$ 为从 P_+ 中独立抽取的随机变量对，其中假设每对不同索引 $i, j \in [n]$ 对应的 X_i和 X'_j 彼此独立。遵循标准设定，一对增强样本通过随机变换

同一个原始样本得到，这一做法在许多实证研究中都有采用（Chen 等人、2020a、b； Dwibedi 等人，2021 年），对于每个 $i \in [n]$，我们考虑 X_i， X_i' 不一定相互独立的情况。经验核对比损失定义如下：

$$\widehat{L}_{\mathrm{KCL}}(f;\lambda) = -\frac{1}{n}\sum_{i=1}^{n} k(f(X_i), f(X_i')) + \frac{\lambda}{n(n-1)}\sum_{i\neq j} k(f(X_i), f(X_j')). \tag{4}$$

在下述表述中，记 $\mathcal{Q} = \{f(\cdot)^\top f(\cdot) : \mathbb{X}\times\mathbb{X} \to \mathbb{R} \mid f \in \mathcal{F}\}$。将拉德马赫复杂度（Mohri 等人，2018年）定义为， $\mathfrak{R}_n^+(\mathcal{Q}) := \mathbb{E}_{P_+,\ \sigma_{1:n}}[\sup_{q\in\mathcal{Q}} n^{-1}\sum_{i=1}^n \sigma_i q(X_i,\ X_i')]$,其中 σ_1， $\cdots$， σ_n 为独立随机变量，且各自以二分之一的概率取值 ± 1。我们还定义拉德马赫复杂度 $\mathfrak{R}_{n/2}^-(\mathcal{Q}; s^*)$，其中从次数为 n 的对称群 S_n 中选取最优的 s^*：

$$\mathfrak{R}_{n/2}^-(\mathcal{Q}; s^*) := \max_{s\in S_n} \mathbb{E}_{\substack{X,X' \\ \sigma_{1:(n/2)}}} \left[\sup_{q\in\mathcal{Q}} \frac{2}{n}\sum_{i=1}^{n/2} \sigma_i q(X_{s(2i-1)}, X'_{s(2i)})\right].$$

注意， $\mathfrak{R}_{n/2}^-(\mathcal{Q};\ s^*)$ 与 针对“独立同分布求和”分块的平均技术相关，该技术用于 U-统计量，并在Cl´emen˙con 等人(2008年中作了解释；更多细节也可参见附录 D.4中的备注 3。KCL 的泛化误差界如下所示：

定理 3。 假设 *1* 成立，且 n 为偶数。此外，假设$\widehat{\ }$ $\widehat{\ }$的极小化器 $f \in \mathcal{F}$ $L_{\mathrm{KCL}}(f;\ \lambda)$ 存在。那么，以至少 $1-2\varepsilon$ 的概率，其中 $\varepsilon > 0$，有

$$L_{\mathrm{KCL}}(\widehat{f};\lambda) \leq L_{\mathrm{KCL}}(f;\lambda) + 2\cdot \mathrm{Gen}(n,\lambda,\varepsilon),$$

其中 $\mathfrak{R}_n(\mathcal{Q})$: $= \mathfrak{R}_n^+(\mathcal{Q}) + \lambda\mathfrak{R}_{n/2}^-(\mathcal{Q};\ s^*)$，并且

$$\mathrm{Gen}(n,\lambda,\varepsilon) = O\left(\mathfrak{R}_n(\mathcal{Q}) + (1+\lambda)\sqrt{\frac{\log(2/\varepsilon)}{n}}\right).$$

注 2。 在附录 *D.2* 中，我们证明，在某些条件下，$\mathrm{Gen}(n,\ \lambda,\ \varepsilon) \downarrow 0$ 成立，且$n \to \infty$。

定理 3 的证明见附录 D.1。由于对于每个 $i \in [n]$， X_i、 X_i' 不一定相互独立，因此标准技术（例如 Mohri 等人（2018）中的定理 3.3）不适用于该证明。我们改用 Zhang 等人（2019）的结果来克服这一困难，这不同于以往针对对比学习的界（Arora 等，2019；Ash 等，2022；HaoChen 等人，2021；Lei 等人，2023；Nozawa 等人，2020；Wang 等人，2022b；Zhang 等，2022；Zou 和 Liu，2023）（更多细节见附录 F.5）。这里，如果 $\mathfrak{R}_n(\mathcal{Q}) \downarrow 0$ 如$n \to \infty$，那么利用我们的结果，我们可以证明，对于每个 $f \in \mathcal{F}$，经验对比损失与总体对比损失一致。

5.4 A理论结果的应用：一种新的替代界

近期的研究（Arora 等人，2019年；Ash 等人，2022年；Awasthi 等人，2022年；Bao 等人，2022年；Dufumier 等人，2022年；HaoChen 等人，2021年；Nozawa 等人，2020年；Nozawa 和 Sato，2021年；Saunshi 等人，2022年；

Tosh 等人，2021b；Wang 等人，2022a；Zou 和 Liu，2023年）表明，某些对比学习目标可以视为下游任务中监督损失或误差的替代损失。这里，Arora 等人（2019）表明，对比损失 L_{CL} 可替代监督损失或误差L_{Sup}：对于每个 $f \in \mathcal{F}$， $L_{\mathrm{sup}}(W \circ f) \lesssim L_{\mathrm{CL}}(f) + \alpha$对某个 $\alpha \in \mathbb{R}$和矩阵 W成立。这类不等式也称为替代界（Bao 等，2022）。Arora 等人（2019）表明，该 $\widehat{}$不等式保证以高概率有 $L_{\mathrm{sup}}(W^* \circ f) \lesssim L_{\mathrm{CL}}(f) + \alpha$成立，其中 f是 L_{CL}的经验损失的极小化器， W^* 是最优权重，而 α 是某个项。受这些研究启发，我们为 KCL 给出了一个替代界。

$\widehat{}$定理 *4*。假设假设 *1* 和假设 *2* 成立， n 为偶函数，并且存在一个极小化器 f $\widehat{}$ $\widehat{}$，使得 $L_{\mathrm{KCL}}(f; \lambda)$ 的 f有意义。取满足假设 *2* 中条件 *(A)*、*(B)* 和 *(C)* 的 $\delta \in \mathbb{R}$、$K \in \mathbb{N}$、 $\mathbb{M}_1$、 $\cdots$、 $\mathbb{M}_K$ 和 y。那么，对于任意 $f \in \mathcal{F}$ 和 $\varepsilon > 0$，以至少 $1 - 2\varepsilon$ 的概率，

$$L_{\mathrm{Err}}(\widehat{f}, W^*, \beta^*; y) \lesssim L_{\mathrm{KCL}}(f; \lambda) + (1 - \frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda\mathfrak{c}(\widehat{f}) + R(\lambda) + 2\mathrm{Gen}(n, \lambda, \varepsilon),$$

$\widehat{}$ $\widehat{}$ $\widehat{}$其中 L_{Err}（f、W^*、β^*；y）=下确界$_{W,\ \beta}$ L_{Err}（f、W、β；y），且 $\lesssim$省略了系数 $8(K-1)/$（$\Delta_{\min}$（f）·最小值$_{i\in[K]}P_{\mathbb{X}}$（$\mathbb{M}_i$））。

定理 4 的证明见附录 E。该定理表明，以高概率最小化 $\mathcal{F}$ 中的核对比损失可以降低分类误差的下确界。请注意，由于假设 2 的条件 (B) 中的关系，较大的 λ 会使 δ 变小；因此，较大的 λ 可能会导致 $(1-\delta/2)\mathfrak{a}(f)-\lambda\mathfrak{c}(f)$ 增大，并使上界 $\widehat{}$ $\widehat{}$变得宽松，前提是 $\mathfrak{a}(f) > 0$且 $\mathfrak{c}(f) < 0$。我们通过实证发现，具有较大 λ的 KCL 框架在下游分类任务中的表现会下降；见附录 H.4。

5.4.1 与相关工作的比较

一些研究也为某些对比学习目标建立了替代界（Arora 等，2019；Ash 等，2022；Awasthi 等人，2022；Bao 等，2022；Dufumier 等，2022；HaoChen 等，2021；Nozawa 等人，2020；Nozawa 和 Sato，2021；Saunshi 等，2022；Tosh 等人，2021b；Wang 等，2022a；Zou 和 Liu，2023）。先前研究与定理 4 之间的主要差异可归纳为三点：1）定理 4 指出，核对比损失是分类误差的一种替代损失，而先前研究处理的是其他对比学习目标。2）近期研究（Bao 等，2022；Wang 等，2022a）证明，InfoNCE 损失是交叉熵损失的一种替代损失。然而，由于分类校准损失理论（例如，见 Zhang（2004））指出，在多分类设置下，分类损失与交叉熵损失之间的关系较为复杂，因此根据已有结果，InfoNCE 损失与分类误差之间的关系并非显而易见。另一方面，结合定理 4 和（1），我们可以证明 InfoNCE 损失也是分类误差的一种替代损失。值得注意的是，定理 4 也适用于其他对比学习目标。3）定理 4 中的界是通过引入第 4 节提出的公式建立的。特别是，我们的界包含几何量 δ 和超参数 λ。

6 结论与讨论

本文研究了对比学习所学习到的表示结构的刻画。通过采用核对比学习（KCL）作为统一框架，我们表明，基于统计相似性的公式能够刻画学习到的表示的簇，并保证最小化核对比损失能够在下游分类任务中取得良好表现。本文的局限性在于，实践中很难计算数据集的真实 sim(x, x'; λ) 和 δ。然而，我们相信，我们的理论将推动未来的理论与实证研究，通过由 sim(x, x'; λ) 和 δ 定义的增强数据集来探究对比学习在实践中取得成功的原因。值得注意的是，近期研究（Tsai 等人，2021，2020）采用神经网络估计逐点依赖，因此这一估计问题是未来一项重要的研究工作。未来值得研究核的选择如何通过我们的理论影响表示的质量。研究 KCL 的迁移学习视角也是一项有趣的未来工作，因为近期研究（HaoChen 等人，2022；Shen 等人，2022；Zhao 等人，2023）也针对一些对比学习框架探讨了这一问题。

伦理声明

由于本文主要研究对比学习的理论分析，我们认为其不会造成直接的负面社会影响。不过，揭示对比学习的详细性质可能会增加相关知识被滥用的风险。我们指出，利用所提出的方法进行此类不当用途并非易事，因为本文并未深入讨论其应用。

参考文献

Arendt, W., Batty, C. J., Hieber, M. 和 Neubrander, F.（2011）。向量值拉普拉斯变换与柯西问题。Birkh¨auser Basel。24 Aronszajn, N.（1950年）。再生核理论。《美国数学学会会刊》，68(3):337–404。3, 5, 51 Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O. 和 Saunshi, N.（2019年）。对比式无监督表征学习的理论分析。载于《第36届国际机器学习会议论文集》，《机器学习研究论文集》第97卷，第5628–5637页。PMLR。2, 4, 8, 9, 11, 12, 13, 14, 40, 47, 48 Ash, J., Goel, S., Krishnamurthy, A. 和 Misra, D.（2022年）。探究负样本在对比表征学习中的作用。载于《第25届国际人工智能与统计学会议论文集》，《机器学习研究论文集》第151卷，第7187–7209页。PMLR。2, 13, 14, 40, 47 Awasthi, P., Dikkala, N. 和 Kamath, P.（2022年）。更多负样本一定会损害对比学习吗？载于《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第1101–1116页。PMLR。2, 13, 14

Bao, H., Nagano, Y. 和 Nozawa, K.（2022年）。对比损失与监督损失之间的替代间隙。载于《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第1585–1606页。PMLR。2, 4, 6, 13, 14 Berlinet, A. 和 Thomas-Agnan, C.（2004年）。概率论与统计学中的再生核希尔伯特空间。Springer Science & Business Media。3, 5, 23, 51 Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. 和 Joulin, A.（2020年）。通过对比聚类分配进行视觉特征的无监督学习。载于《神经信息处理系统进展》，第33卷，第9912–9924页。Curran Associates, Inc.。1, 2, 3 Chen, T., Kornblith, S., Norouzi, M. 和 Hinton, G.（2020a）。用于视觉表征对比学习的简单框架。载于《国际机器学习会议论文集》，第1597–1607页。PMLR。1, 2, 3, 4, 5, 6, 11, 12, 13, 44, 55, 57, 58 Chen, T., Luo, C. 和 Li, L.（2021年）。对比损失的有趣特性。载于《神经信息处理系统进展》，第34卷，第11834–11845页。Curran Associates, Inc.。3, 4, 6, 7, 44, 60 Chen, X., Fan, H., Girshick, R. 和 He, K.（2020b）。基于动量对比学习的改进基线。arXiv预印本 arXiv:2003.04297v1。1, 2, 3, 4, 12, 13, 56 Chen, X. 和 He, K.（2021年）。探索简单的孪生表征学习。载于《2021年IEEE/CVF计算机视觉与模式识别会议（CVPR）》，第15745–15753页。3, 11, 55, 56, 57, 60 Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A. 和 Jegelka, S.（2020年）。去偏对比学习。载于《神经信息处理系统进展》，第33卷，第8765–8775页。Curran Associates, Inc.。4 Cl´emen¸con, S., Lugosi, G. 和 Vayatis, N.（2008年）。U统计量的排序与经验最小化。《统计年鉴》，36(2):844–874。13, 39, 40, 48 Coates, A., Ng, A. 和 Lee, H.（2011年）。无监督特征学习中单层网络的分析。载于《第十四届国际人工智能与统计学会议论文集》，《机器学习研究论文集》第15卷，第215–223页。PMLR。57 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. 和 Fei-Fei, L.（2009年）。ImageNet：大型层次图像数据库。载于《2009年IEEE计算机视觉与模式识别会议》，第248–255页。Ieee。57 Dubois, Y., Ermon, S., Hashimoto, T. 和 Liang, P.（2022年）。通过刻画理想化表征改进自监督学习。载于《神经信息处理系统进展》。4 Dufumier, B., Barbano, C. A., Louiset, R., Duchesnay, E. 和 Gori, P.（2022年）。重新思考使用核进行对比学习的正样本采样。arXiv预印本 arXiv:2206.01646v1。2, 4, 8, 9, 13, 14

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P. 和 Zisserman, A.（2021年）。借助朋友的少许帮助：视觉表征的最近邻对比学习。载于 *2021*年*IEEE/CVF*国际计算机视觉大会（*ICCV*），第9568–9577页。 1、2、3、4、5、6、12、13、55

Gao, T.、Yao, X. 和 Chen, D.（2021年）。SimCSE：句子嵌入的简单对比学习。载于*2021*年自然语言处理经验方法会议论文集，第6894–6910页。计算语言学协会。 3

Goyal, P.、Doll´ar, P.、Girshick, R.、Noordhuis, P.、Wesolowski, L.、Kyrola, A.、Tulloch, A.、Jia, Y. 和 He, K.（2017年）。准确的大批量随机梯度下降：在1小时内训练ImageNet。 *arXiv*预印本 *arXiv:1706.02677v2*。56

Gretton, A., Bousquet, O., Smola, A., and Sch¨olkopf, B.（2005年）。使用希尔伯特–施密特范数测量统计依赖性。载于 算法学习理论，第 63–77 页。Springer Berlin Heidelberg。 4

Grill, J.-B., Strub, F., Altch´e, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M.（2020）。自举潜在表征——一种自监督学习的新方法。载于《神经信息处理系统进展》，第33卷，第21271–21284页。Curran Associates, Inc. 3, 11 Guedj, B.（2019）。PAC贝叶斯学习入门。arXiv预印本 arXiv:1901.05353v3。47 HaoChen, J. Z. and Ma, T.（2023）。对比学习中归纳偏置的理论研究。第十一届国际学习表征会议。https://openreview. net/forum?id=AuEgNlEAmed。2, 4, 8, 9, 45, 46 HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T.（2021）。基于谱对比损失的自监督深度学习的可证明保证。载于《神经信息处理系统进展》，第34卷，第5000–5011页。Curran Associates, Inc. 1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 14, 45, 47, 54, 56, 57, 58 HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T.（2022）。超越可分性：分析对比表征向相关子群体的线性可迁移性。载于《神经信息处理系统进展》。4, 15 Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R´ıo, J. F., Wiebe, M., Peterson, P., G´erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.（2020）。使用 NumPy 进行数组编程。《自然》，585(7825):357–362。58 He, K., Zhang, X., Ren, S., and Sun, J.（2016）。用于图像识别的深度残差学习。载于2016年IEEE计算机视觉与模式识别会议（CVPR），第770–778页。5, 56, 57

Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., and Zhao, H.（2021）。关于自监督学习中的特征去相关。载于2021年IEEE/CVF国际计算机视觉大会（ICCV），第9578–9588页。11 Huang, W., Yi, M., Zhao, X., and Jiang, Z.（2023）。迈向对比自监督学习的泛化。第十一届国际学习表征会议。https://openreview.net/forum?id=XDJwuEYHhme。2, 3, 4, 7, 9, 10, 11, 12, 48, 49 Hunter, J. D.（2007）。Matplotlib：一个二维图形环境。《计算科学与工程》，9(3):90–95。58 Ioffe, S. and Szegedy, C.（2015）。批归一化：通过减少内部协变量偏移加速深度网络训练。载于国际机器学习会议，第448–456页。PMLR。55 Jiang, D., Li, W., Cao, M., Zou, W., and Li, X.（2021）。语音 SimCLR：结合对比目标与重构目标进行自监督语音表征学习。载于Interspeech 2021会议论文集，第1544–1548页。3 Jing, L., Vincent, P., LeCun, Y., and Tian, Y.（2022）。理解对比自监督学习中的维度坍塌。国际学习表征会议。https://openreview.net/forum?id=YevsQ05DEN7。11 Johnson, D. D., Hanchi, A. E., and Maddison, C. J.（2023）。对比学习能够为近似视图不变函数找到最优基。第十一届国际学习表征会议。https://openreview.net/forum?id=AjC0KBjiMu。4, 7, 8 Kiani, B. T., Balestriero, R., Chen, Y., Lloyd, S., and LeCun, Y.（2022）。核机制中的联合嵌入式自监督学习。arXiv预印本arXiv:2209.14884v1。4 Krizhevsky, A.（2009）。从微型图像中学习多层特征。技术报告。57 Lei, Y., Yang, T., Ying, Y., and Zhou, D.-X.（2023）。对比表征学习的泛化分析。arXiv预印本arXiv:2302.12383v2。13, 47 Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A.（2021）。通过核依赖最大化进行自监督学习。载于《神经信息处理系统进展》，第34卷，第15543–15556页。Curran Associates, Inc. 2, 3, 4, 6, 11, 43, 44 Loshchilov, I. and Hutter, F.（2017）。随机梯度下降重启：带热重启的随机梯度下降。国际学习表征会议。https://openreview.net/forum?id=Skq89Scxx。56 McDiarmid, C.（1989）。关于有界差分方法，第148–188页。伦敦数学会讲义系列。剑桥大学出版社。37 Mohri, M., Rostamizadeh, A., and Talwalkar, A.（2018）。机器学习基础。麻省理工学院出版社。13, 33, 38, 47

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Sch¨olkopf, B.（2017）。分布的核均值嵌入：综述与进展。《机器学习基础与趋势》，10(1- 2):1–141。6, 23 Ng, A. Y., Jordan, M. I., and Weiss, Y.（2002）。论谱聚类：分析与算法。载于《神经信息处理系统进展》，第849–856页。4 Nozawa, K., Germain, P., and Guedj, B.（2020）。PAC-贝叶斯对比式无监督表征学习。载于人工智能不确定性会议，第21–30页。PMLR。13, 14, 47 Nozawa, K. and Sato, I.（2021）。理解实例判别式自监督表征学习中的负样本。载于《神经信息处理系统进展》，第34卷，第5784–5797页。Curran Associates, Inc. 13, 14 Parulekar, A., Collins, L., Shanmugam, K., Mokhtari, A., and Shakkottai, S.（2023）。InfoNCE 损失可证明能够学习保持聚类结构的表征。arXiv预印本 arXiv:2302.07920v1。9 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.（2019）。PyTorch：一种命令式、高性能深度学习库。载于《神经信息处理系统进展》，第32卷。Curran Associates, Inc. 55, 58 Poole, B., Ozair, S., van den Oord, A., Alemi, A., and Tucker, G.（2019）。论互信息的变分界。载于《国际机器学习会议》，第5171–5180页。PMLR。4 Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., and Sra, S.（2021a）。对比学习能否避免捷径解？载于《神经信息处理系统进展》，第34卷，第4974–4986页。Curran Associates, Inc. 3, 4 Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S.（2021b）。使用困难负样本的对比学习。《国际学习表征会议》。https://openreview. net/forum?id=CR1XOQ0UTh-。3, 4, 11, 12, 48 Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A.（2022）。理解对比学习需要纳入归纳偏置。载于《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第19250–19286页。PMLR。4, 8, 11, 13, 14 Shalev-Shwartz, S. and Ben-David, S.（2014）。理解机器学习：从理论到算法。剑桥大学出版社。33, 39 Shen, K., Jones, R. M., Kumar, A., Xie, S. M., Haochen, J. Z., Ma, T., and Liang, P.（2022）。建立连接，而非坍塌：解释用于无监督领域自适应的对比学习。载于《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第19847–19878页。PMLR。4, 15

Shi, J. and Malik, J.（2000）。归一化割与图像分割。IEEE模式分析与机器智能汇刊, 22(8):888–905。49, 50 Singh, A.（2021）。Clda：用于半监督领域自适应的对比学习。载于《神经信息处理系统进展》，第34卷，第5089–5101页。Curran Associates, Inc. 3 Smola, A., Gretton, A., Song, L., and Sch¨olkopf, B.（2007年）。分布的希尔伯特空间嵌入。载于《算法学习理论》，第13–31页。Springer Berlin Heidelberg。23, 43 Steinwart, I. and Christmann, A.（2008年）。支持向量机。Springer Science & Business Media。3, 5, 51 Susmelj, I., Helle, M., Wirth, P., Prescott, J., and Ebner et al., M.（2020）。Lightly。https: // github.com/lightly-ai/lightly。57, 58 Terada, Y. and Yamamoto, M.（2019年）。核归一化割：理论再探。载于《国际机器学习会议》，第6206–6214页。PMLR。8, 49, 50 Tian, Y.（2022）。通过逐坐标优化理解深度对比学习。载于《神经信息处理系统进展》。4, 8, 54 Tian, Y., Krishnan, D., and Isola, P.（2020）。对比式多视图编码。载于《计算机视觉——ECCV 2020》，第776–794页。Springer International Publishing。57 TorchVision 维护者和贡献者（2016）。TorchVision：PyTorch 的计算机视觉库。https://github.com/pytorch/vision。58 Tosh, C., Krishnamurthy, A., and Hsu, D.（2021a）。对比估计向线性模型揭示主题后验信息。《机器学习研究期刊》，22(281):1–31。8 Tosh, C., Krishnamurthy, A., and Hsu, D.（2021b）。对比学习、多视图冗余与线性模型。载于《第32届国际算法学习理论大会论文集》，《机器学习研究论文集》第132卷，第1179–1206页。PMLR。4, 7, 14 Trillos, N. G., Slepˇcev, D., Von Brecht, J., Laurent, T., and Bresson, X.（2016）。Cheeger割与比率图割的一致性。《机器学习研究期刊》, 17(1):6268–6313。50 Tsai, Y.-H. H., Li, T., Ma, M. Q., Zhao, H., Zhang, K., Morency, L.-P., and Salakhutdinov, R.（2022）。基于核的条件对比学习。《国际学习表征会议》。https://openreview.net/forum?id=AAJLBoGt0XM。4 Tsai, Y.-H. H., Ma, M. Q., Yang, M., Zhao, H., Morency, L.-P., and Salakhutdinov, R.（2021）。采用相对预测编码的自监督表征学习。《国际学习表征会议》。https://openreview.net/forum?id=068E_JSq9O。15 Tsai, Y.-H. H., Zhao, H., Yamada, M., Morency, L.-P., and Salakhutdinov, R. R.（2020）。逐点依赖关系估计的神经方法。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. 编，《神经信息处理系统进展》，第33卷，第62–72页。Curran Associates, Inc. 7, 15

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., 和 Lucic, M.（2020年）。关于表征学习中的互信息最大化。国际学习表征会议。https://openreview.net/forum?id=rkxoh24FPH。4 Tu, Z., Zhang, J., 和 Tao, D.（2019年）。对抗学习的理论分析：一种极小极大方法。载于《神经信息处理系统进展》，第32卷。Curran Associates, Inc. 35, 36 van den Oord, A., Li, Y., 和 Vinyals, O.（2018年）。基于对比预测编码的表征学习。arXiv预印本 arXiv:1807.03748v2。3, 6, 44 Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., 和 Van Gool, L.（2020年）。Scan：学习在没有标签的情况下对图像进行分类。载于《欧洲计算机视觉会议论文集》。3 von K¨ugelgen, J., Sharma, Y., Gresele, L., Brendel, W., Sch¨olkopf, B., Besserve, M., 和 Locatello, F.（2021年）。通过数据增强实现自监督学习，且可证明能将内容与风格分离。载于《神经信息处理系统进展》，第34卷，第16451–16467页。Curran Associates, Inc. 4, 9 Von Luxburg, U.（2007年）。谱聚类教程。《统计与计算》，17(4):395–416。49, 50, 54 Wainwright, M. J.（2019年）。《高维统计：非渐近视角》。剑桥统计与概率数学丛书。剑桥大学出版社。34, 35, 36, 39 Wang, F. 和 Liu, H.（2021年）。理解对比损失的行为。载于《2021年IEEE/CVF计算机视觉与模式识别会议（CVPR）》，第2495–2504页。7 Wang, T. 和 Isola, P.（2020年）。通过超球面上的对齐与均匀性理解对比表征学习。载于《国际机器学习会议》，第9929–9939页。PMLR。4, 5, 6, 7, 10, 44 Wang, Y., Zhang, Q., Wang, Y., Yang, J., 和 Lin, Z.（2022a）。混沌是一架梯子：通过增强重叠对对比学习的新理论理解。国际学习表征会议。https://openreview.net/forum?id=ECvgmYVyeUz。2, 4, 5, 8, 9, 11, 14 Wang, Z., Luo, Y., Li, Y., Zhu, J., 和 Sch¨olkopf, B.（2022b）。条件矩模型的谱表征学习。arXiv预印本 arXiv:2210.16525v2。7, 8, 13, 47, 48 Waskom, M. L.（2021年）。seaborn：统计数据可视化。《开放源代码软件期刊》，6(60):3021。58 Wen, Z. 和 Li, Y.（2021年）。迈向理解自监督对比学习的特征学习过程。载于《第38届国际机器学习会议论文集》，《机器学习研究论文集》第139卷，第11112–11122页。PMLR。4 Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y., 和 LeCun, Y.（2022年）。解耦对比学习。载于《计算机视觉——ECCV 2022》，第668–684页。施普林格·自然瑞士。7, 44

Zhang, G., Lu, Y., Sun, S., Guo, H., 和 Yu, Y.（2022年）。f-互信息对比学习。https://openreview.net/forum?id=3kTt_W1_tgw。4, 13, 47 Zhang, R. R., Liu, X., Wang, Y., 和 Wang, L.（2019年）。图相关变量的 McDiarmid 型不等式与稳定性界。载于《神经信息处理系统进展》，第32卷。Curran Associates, Inc. 13, 36, 37, 38, 47, 48 Zhang, T.（2004年）。若干多类别大间隔分类方法的统计分析。《机器学习研究期刊》，5:1225–1251。14 Zhao, X., Du, T., Wang, Y., Yao, J., 和 Huang, W.（2023年）。ArCL：通过增强鲁棒表示提升对比学习。第十一届国际学习表征会议。https://openreview.net/forum?id=n0Pb9T5kmb。2, 3, 4, 9, 12, 15, 48, 49 Zou, X. 和 Liu, W.（2023年）。对抗对比学习的泛化界。arXiv预印本 arXiv:2302.10633v1。2, 13, 14, 47

第 2.2 节中的证明

首先证明命题 1。

命题 *1* 的证明。 该结论的证明与 Muandet 等人（2017年）中的引理 3.1 的证明（另见 Smola 等人（2007年））基本一致。该证明表明，若 $\mathbb{E}_P[\sqrt{k(x,\ x)}] < +\infty$，其中 $x \sim P$，则 $\mathbb{E}_P[k(\cdot,\ x)] \in \mathcal{H}_k$。为完整起见，我们稍作修改 Muandet 等人（2017年）的证明，以给出命题 1 的证明。

设 $\mathbb{M}$ 是 $\mathbb{X}$中的一个可测集。定义 $\mu_{\mathbb{M}}(f) := \mathbb{E}[h(f(x))|\mathbb{M}]$。我们的目标是证明$\mu_{\mathbb{M}}(f) \in \mathcal{H}_k$ 成立。为此，对于 $\phi \in \mathcal{H}_k$，我们计算

$$
\begin{aligned}
|\mathbb{E}\left[\phi(f(x))|\mathbb{M}\right]| &= |\mathbb{E}\left[\langle \phi, k(\cdot, f(x))\rangle_{\mathcal{H}_k}|\mathbb{M}\right]| \\
&\leq \mathbb{E}\left[|\langle \phi, k(\cdot, f(x))\rangle_{\mathcal{H}_k}|\,|\mathbb{M}\right] \\
&\leq \mathbb{E}\left[\|\phi\|_{\mathcal{H}_k}\|k(\cdot, f(x))\|_{\mathcal{H}_k}|\mathbb{M}\right] \qquad \text{（柯西-施瓦茨不等式）} \\
&= \|\phi\|_{\mathcal{H}_k}\mathbb{E}\left[\sqrt{k(f(x), f(x))}|\mathbb{M}\right].
\end{aligned}
$$

由于 $\sup_{z,z' \in \mathbb{S}^{d-1}} k(z, z') < \infty$成立，我们有 $\mathbb{E}[\sqrt{k(f(x)\ f(x))|\mathbb{M}}] < +\infty$。因此，映射 $\phi \mapsto \mathbb{E}[\phi(f(x))|\mathbb{M}]$ 是 $\mathcal{H}_k$上的有界线性泛函。由里斯表示定理，存在某个 $\xi \in \mathcal{H}_k$，使得 $\mathbb{E}[\phi(f(x))|\mathbb{M}] = \langle \xi, \phi\rangle_{\mathcal{H}_k}$。 然而，令 $\phi = k(\cdot, z)$，则 $\xi(z) = \langle \xi, k(\cdot, z)\rangle_{\mathcal{H}_k} = \mathbb{E}[k(f(x)\ z)|\mathbb{M}]$。这意味着 $\xi = \mathbb{E}[k(f(x)\ \cdot)|\mathbb{M}] \in \mathcal{H}_k$。 由于 k 是对称的，我们有 $\mu_{\mathbb{M}}(f) = \mathbb{E}[k(\cdot, f(x))|\mathbb{M}] \in \mathcal{H}_k$。 □

第 B 节 5.1 节中的证明

B.1 定理 1 证明中用到的引理

在证明定理 1 之前，我们先给出若干证明中会用到的基本且有用的引理。由于定义 $\mu_{\mathbb{M}}(f)$，其中 $\mathbb{M}$是 $\mathbb{X}$的可测子集，且 $f \in \mathcal{F}$，与通常形式的核均值嵌入（Berlinet andThomas-Agnan，2004年；Muandet 等人，2017年）略有不同，这是由编码器函数 f的存在所致，因此为求完整，我们给出每个引理的证明。

引理 1. 设 $\mathcal{H}_k$ 的一组正交归一基为 $\{e_j\}$，且 $\mathbb{M}$ 为一个可测集。设 $f \in \mathcal{F}$。则对于每个 j，以下恒等式成立：

$$
\int_{\mathbb{M}} \langle h(f(x)), e_j\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) = \langle \mu_{\mathbb{M}}(f), e_j\rangle_{\mathcal{H}_k}.
$$

证明. 计算得,

$$\begin{aligned}
\langle \mu_{\mathbb{M}}(f), e_j \rangle_{\mathcal{H}_k} &= \left\langle \int_{\mathbb{M}} h(f(x)) P_{\mathbb{X}}(dx|\mathbb{M}), e_j \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \int_{\mathbb{M}} \sum_{j'} \langle h(f(x)), e_{j'} \rangle_{\mathcal{H}_k} e_{j'} P_{\mathbb{X}}(dx|\mathbb{M}), e_j \right\rangle_{\mathcal{H}_k} \\
&= \left\langle \sum_{j'} e_{j'} \int_{\mathbb{M}} \langle h(f(x)), e_{j'} \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}), e_j \right\rangle_{\mathcal{H}_k} \\
&= \int_{\mathbb{M}} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}),
\end{aligned}$$

其中，在第三行中，我们对 Bochner 积分使用了控制收敛定理（例如，参见 Arendt 等人 （2011））。因此，我们得到了该结论。 □

引理 2。 设 $\mathbb{M}$ 和 $\mathbb{M}'$ 是 $\mathbb{X}$ 的可测子集。设 $f \in \mathcal{F}$。则有

$$\int_{\mathbb{M}} \int_{\mathbb{M}'} \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') = \langle \mu_{\mathbb{M}}(f), \mu_{\mathbb{M}'}(f) \rangle_{\mathcal{H}_k}.$$

证明。 设 $\{e_j\}$ 是 $\mathcal{H}_k$ 的正交归一基。则有,

$$\begin{aligned}
&\int_{\mathbb{M}} \int_{\mathbb{M}'} \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \\
&= \int_{\mathbb{M}} \int_{\mathbb{M}'} \left\langle \sum_j \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} e_j, \sum_j \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} e_j \right\rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \\
&= \int_{\mathbb{M}} \int_{\mathbb{M}'} \sum_j \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \\
&= \sum_j \int_{\mathbb{M}} \int_{\mathbb{M}'} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) P_{\mathbb{X}}(dx'|\mathbb{M}') \qquad (5) \\
&= \sum_j \left(\int_{\mathbb{M}} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx|\mathbb{M}) \right) \left(\int_{\mathbb{M}'} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} P_{\mathbb{X}}(dx'|\mathbb{M}') \right) \\
&= \sum_j \langle \mu_{\mathbb{M}}(f), e_j \rangle_{\mathcal{H}_k} \langle \mu_{\mathbb{M}'}(f), e_j \rangle_{\mathcal{H}_k} \qquad \text{(Lemma 1)} \\
&= \left\langle \sum_j \langle \mu_{\mathbb{M}}(f), e_j \rangle_{\mathcal{H}_k} e_j, \sum_j \langle \mu_{\mathbb{M}'}(f), e_j \rangle_{\mathcal{H}_k} e_j \right\rangle_{\mathcal{H}_k} \\
&= \langle \mu_{\mathbb{M}}(f), \mu_{\mathbb{M}'}(f) \rangle_{\mathcal{H}_k},
\end{aligned}$$

其中 （5）我们使用控制收敛定理。因此得到所述结论。 □

定理 1 的证明

定理1的证明中使用以下符号。

定义 3。 记为 $M_k = \sup_{z,\ z' \in \mathbb{S}^{d-1}} \|k(\cdot, z) - k(\cdot, z')\|^2_{\mathcal{H}_k}$。我们定义：

$$R(\lambda) := \frac{M_k}{2} \sum_{i \neq j} P_+((\mathbb{M}_i \cap \mathbb{M}_j) \times (\mathbb{M}_i \cap \mathbb{M}_j)) + \lambda\psi(1) \sum_{i=1}^{K} P_{\mathbb{X}}(\mathbb{M}_i)\,(1 - P_{\mathbb{X}}(\mathbb{M}_i)) + (1-\lambda)\psi(1),$$

其中 $dP_+(x, x') = w(x, x')d\nu^{\otimes 2}_{\mathbb{X}}(x, x')$。

注意，在假设 1下， $k(z) := k(z,\ z) = \psi(z^\top z) = \psi(1)$ 在 $\mathbb{S}^{d-1}$上是一个常值函数。现在我们可以给出定理 1 的证明。

定理 *1* 的证明。 分析以下形式比分析核对比损失更方便：

$$\widetilde{L}_{\mathrm{KCL}}(f; \lambda) := \underbrace{\mathbb{E}_{x,x^+}\left[\|h(f(x)) - h(f(x^+))\|^2_{\mathcal{H}_k}\right]}_{\text{the positive term}} - \lambda \underbrace{\mathbb{E}_{x,x^-}\left[\|h(f(x)) - h(f(x^-))\|^2_{\mathcal{H}_k}\right]}_{\text{the negative term}}. \tag{6}$$

注意， $\widetilde{L}_{\mathrm{KCL}}(f;\ \lambda) = 2(1-\lambda)\psi(1) + 2L_{\mathrm{KCL}}(f;\ \lambda)$ 成立，因为对于所有 $x \in \mathbb{X}$， $f(x) \in \mathbb{S}^{d-1}$。对于 $L_{\mathrm{KCL}}(f)$ 的正项，我们可以计算得：

$$\begin{aligned}
&\mathbb{E}_{x,x^+}\left[\|h(f(x)) - h(f(x^+))\|^2_{\mathcal{H}_k}\right] \\
&\geq \int_{\bigcup_{i=1}^K \mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x, x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \qquad (7) \\
&\geq \sum_{i=1}^{K} \int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x, x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&\quad - \sum_{j \neq i} \int_{(\mathbb{M}_i \cap \mathbb{M}_j) \times (\mathbb{M}_i \cap \mathbb{M}_j)} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x, x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \qquad (8) \\
&\geq \sum_{i=1}^{K} \left(\int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x, x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') - M_k \sum_{j \neq i} P_+((\mathbb{M}_i \cap \mathbb{M}_j) \times (\mathbb{M}_i \cap \mathbb{M}_j)) \right) \qquad (9)
\end{aligned}$$

其中，在第二个不等式中，我们利用了以下事实：

$$Q(\bigcup_{i=1}^{K} \mathbb{M}_i \times \mathbb{M}_i) \geq \sum_{i=1}^{K} Q(\mathbb{M}_i \times \mathbb{M}_i) - \sum_{i \neq j} Q((\mathbb{M}_i \times \mathbb{M}_i) \cap (\mathbb{M}_j \times \mathbb{M}_j)),$$

对于任意概率测度 Q in $\mathbb{X} \times \mathbb{X}$，在最后一个不等式中，我们利用定义 $M_k = \sup_{z,\ z' \in \mathbb{S}^{d-1}} \|k(\cdot, z) - k(\cdot, z')\|^2_{\mathcal{H}_k}$。上述下界中的第一项可界定为

$$\begin{aligned}
&\sum_{i=1}^{K} \int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x,x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
\geq &\sum_{i=1}^{K} \int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} \cdot (\lambda + \delta) w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x'), \qquad (10)
\end{aligned}$$

其中我们利用 $\mathbb{M}_i$ 对每个 $i \in \{1, \cdots, K\}$ 的定义；回想一下，由于假设 **(B)2** 中的条件 2，对于每个 $x, x' \in \mathbb{M}_i$ 我们有 $\mathsf{sim}(x,\ x';\ \lambda) \geq \delta$。

另一方面，对于负项，我们可以按如下方式计算：

$$\begin{aligned}
&- \mathbb{E}_{x,x^-} \left[\|h(f(x)) - h(f(x^-))\|^2_{\mathcal{H}_k} \right] \\
= &- \int_{\bigcup_{i=1}^{K} \mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&- \int_{\mathbb{X} \times \mathbb{X} \setminus (\bigcup_{i=1}^{K} \mathbb{M}_i \times \mathbb{M}_i)} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
\geq &- \sum_{i=1}^{K} \int_{\mathbb{M}_i \times \mathbb{M}_i} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&- \int_{\mathbb{X} \times \mathbb{X} \setminus (\bigcup_{i=1}^{K} \mathbb{M}_i \times \mathbb{M}_i)} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x'), \qquad (11)
\end{aligned}$$

其中，最后一个不等式是由并集界得到的。对于上述不等式右侧的第二项，我们有

$$\begin{aligned}
&- \int_{\mathbb{X} \times \mathbb{X} \setminus (\bigcup_{i=1}^{K} \mathbb{M}_i \times \mathbb{M}_i)} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
\geq &- \sum_{i \neq j} \int_{\mathbb{M}_i \times \mathbb{M}_j} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&- \int_{\mathbb{X} \times \mathbb{X} \setminus (\bigcup_{i,j=1}^{K} \mathbb{M}_i \times \mathbb{M}_j)} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \qquad (12) \\
\geq &- \sum_{i \neq j} \int_{\mathbb{M}_i \times \mathbb{M}_j} \|h(f(x)) - h(f(x'))\|^2_{\mathcal{H}_k} w(x) w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') - M_k P_{\mathbb{X}}^{\otimes 2} \left(\mathbb{X} \times \mathbb{X} \setminus \left(\bigcup_{i,j=1}^{K} \mathbb{M}_i \times \mathbb{M}_j \right) \right). \qquad (13)
\end{aligned}$$

这里，第二个 t （13）中的项为零，因为假设2 意味着 $\mathbb{X} \times \mathbb{X} = \bigcup_{i,\ j=1}^{K} \mathbb{M}_i \times \mathbb{M}_j$。Th e

（13）中的第一项还可进一步下界为：

$$
\begin{aligned}
&-\sum_{i\neq j}\int_{\mathbb{M}_i\times\mathbb{M}_j}\|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')\\
\geq\ & 2\sum_{i\neq j}\Bigg\{-\sup_{z\in\mathbb{S}^{d-1}}k(z,z)P_{\mathbb{X}}(\mathbb{M}_i)P_{\mathbb{X}}(\mathbb{M}_j)\\
&\qquad+\int_{\mathbb{M}_i\times\mathbb{M}_j}\langle h(f(x)),h(f(x'))\rangle_{\mathcal{H}_k}w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')\Bigg\} \qquad (14)\\
=\ & 2\sum_{i\neq j}\left\{-\sup_{z\in\mathbb{S}^{d-1}}k(z,z)P_{\mathbb{X}}(\mathbb{M}_i)P_{\mathbb{X}}(\mathbb{M}_j)+P_{\mathbb{X}}(\mathbb{M}_i)P_{\mathbb{X}}(\mathbb{M}_j)\langle\mu_i(f),\mu_j(f)\rangle_{\mathcal{H}_k}\right\} \qquad \text{(Lemma 2)}\\
=\ & -2\sup_{z\in\mathbb{S}^{d-1}}k(z,z)\sum_{i=1}^{K}P_{\mathbb{X}}(\mathbb{M}_i)\left(1-P_{\mathbb{X}}(\mathbb{M}_i)\right)+2\mathfrak{c}(f).
\end{aligned}
$$

因此，对于负项，我们得到如下不等式，

$$
\begin{aligned}
&-\mathbb{E}_{x,x^-}\left[\|h(f(x))-h(f(x^-))\|_{\mathcal{H}_k}^2\right]\\
\geq\ & -\sum_{i=1}^{K}\int_{\mathbb{M}_i\times\mathbb{M}_i}\|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x)w(x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')\\
&+2\mathfrak{c}(f)-2\sup_{z\in\mathbb{S}^{d-1}}k(z,z)\sum_{i=1}^{K}P_{\mathbb{X}}(\mathbb{M}_i)\left(1-P_{\mathbb{X}}(\mathbb{M}_i)\right). \qquad (15)
\end{aligned}
$$

结合（6）、（9）、（10）和（15），可得

$$
\begin{aligned}
&L_{\mathrm{KCL}}(f;\lambda)+(1-\lambda)\psi(1)\\
\geq\ & \frac{\delta}{2}\sum_{i=1}^{K}\mathbb{E}_{x,x^-}\left[\|h(f(x))-h(f(x^-))\|_{\mathcal{H}_k}^2;\mathbb{M}_i\times\mathbb{M}_i\right]+\lambda\mathfrak{c}(f)\\
&-\frac{M_k}{2}\sum_{i\neq j}P_+((\mathbb{M}_i\cap\mathbb{M}_j)\times(\mathbb{M}_i\cap\mathbb{M}_j))-\lambda\psi(1)\sum_{i=1}^{K}P_{\mathbb{X}}(\mathbb{M}_i)\left(1-P_{\mathbb{X}}(\mathbb{M}_i)\right). \qquad (16)
\end{aligned}
$$

至此，证明完成。 □

B.3 推论1的证明

推论*1*的证明。推论1的证明只需检查定理1证明中出现的每个不等式是否都取等号。下面列出具体的检查过程：

(7)：由于对于任意 $(x,x')\in\mathbb{M}_i\times\mathbb{M}_j(i\neq j)$，$w(x,x')=0$，我们有 $\int_{\mathbb{M}_i\times\mathbb{M}_j}\|h(f(x))-h(f(x'))\|_{\mathcal{H}_k}^2 w(x,x')d\nu_{\mathbb{X}}(x)d\nu_{\mathbb{X}}(x')=0$。这里，我们有分解 $\mathbb{X}\times\mathbb{X}=(\bigcup_{i=1}^{K}\mathbb{M}_i)\times(\bigcup_{j=1}^{K}\mathbb{M}_j)=\bigcup_{i,j=1}^{K}\mathbb{M}_i\times\mathbb{M}_j$，其中对于任意 (i,j,i',j')，$(\mathbb{M}_i\times\mathbb{M}_j)\cap(\mathbb{M}_{i'}\times\mathbb{M}_{j'})=\emptyset$

使得 $i \neq i'$ 或 $j \neq j'$；根据 $\mathbb{M}_1, \cdots, \mathbb{M}_K$ 两两不相交这一假设，利用概率测度的可加性即可得到该等式。

(8)：由于 $\mathbb{M}_i \cap \mathbb{M}_j = \emptyset$ 对于 $i, j \in [K]$ 使得 $i \neq j$，右侧（8）中的第一项等于左侧（8）中的第一项。另一方面，右侧（8）中的第二项等于 0。因此，等式成立。

(9)：由于在 $\mathbb{M}_i \cap \mathbb{M}_j = \emptyset$ 对于 i 时，基于 j $(i \neq j)$ 这一假设，（

(10)：根据以下假设，等式成立：对于任意 $x, x' \in \mathbb{M}_i (i \in [K])$，$\mathsf{sim}(x, x'; \lambda) = \delta$ 成立。事实上，这一假设意味着 $w(x, x') = (\lambda + \delta) w(x) w(x')$ 对于任意 $x, x' \in \mathbb{M}_i$。

(11)：由于 $\mathbb{M}_1 \times \mathbb{M}_1, \cdots, \mathbb{M}_K \times \mathbb{M}_K$ 彼此不相交，因此等式成立。

(12)：由于 $\bigcup_{i,j=1}^{K} \mathbb{M}_i \times \mathbb{M}_j = \mathbb{X} \times \mathbb{X}$，(12) 右端的第二项等于 0。因此，等式成立。

(13)：由于与上文 (12) 相同的原因，等式成立。

(14)：由于 $\|h(f(x))\|_{\mathcal{H}_k}^2 = k(f(x)\ f(x)) = \psi(f(x)^\top f(x)) = \psi(1)$，对于任意 $x \in \mathbb{X}$ 和 $f \in \mathcal{F}$，等式成立。

(15)：由于 (15) 是由 (11)、(12)、(13) 和 (14) 组合而成，因此在此情况下，等式 (15) 成立。

(16)：由于 (16) 是通过合并 (9)、(10) 和 (15) 得到的，因此等式成立。

因此，我们得到了该结果。 □

C 第 5.2 节证明

我们给出定理 2 的一个推广版本的证明。推广后的定理如下所示。

定理 5（定理2的推广）。假设假设 *1* 和 *2* 成立。取 $K \in \mathbb{N}$ 和 $\mathbb{M}_1, \cdots, \mathbb{M}_K$ 使得假设 **（A）**中的条件 *2* 得到满足。令，为 $\mathbb{X}$ 的一个不交划分，且对每个 $\mathbb{M}_i \subset \mathbb{M}_i$ $i \in [K]$ 均满足。定义：$\widetilde{y}\ \mathbb{X} \to [K]$ 为（）对每个 $x \in \mathbb{M}_i$ 成立。于是，对于每个有意义的编码器 $f \in \mathcal{F}$，有

$$L_{\mathrm{Err}}(f, W_\mu, \beta_\mu; \widetilde{y}) \leq \frac{8(K-1)}{\Delta_{\min}(f) \cdot \min_{i \in [K]} P_{\mathbb{X}}(\mathbb{M}_i)} \mathfrak{a}(f)$$

其中 $\Delta_{\min}(f)$ =最小值$_{i \neq j} \|\mu_i(f) - \mu_j(f)\|_{\mathcal{H}_k}^2$。

定理5的证明。根据定义，我们有 $\mathbb{M}_i \subset \mathbb{M}_i$ 且 $\mathbb{M}_i \cap \mathbb{M}_j = \emptyset$ 对于所有互不相同的指标对 $i, j \in [K]$。回顾一下 $L_{\mathrm{Err}}(f, W_\mu, \beta_\mu; \widetilde{y})$ 的定义：

$$L_{\mathrm{err}}(f, W_\mu, \beta_\mu; \widetilde{y}) = P_{\mathbb{X}}\left(g_{f, W_\mu, \beta_\mu}(x) \neq \widetilde{y}(x)\right).$$

这里回顾一下，我们也允许 arg max、arg min 在出现并列时任意打破平局。例如，如果存在互不相同的整数 i_1、$\cdots$、$i_j \in [K]$，使得 g_f、w_μ、$\beta_\mu(x) = \{i_1, \cdots, i_j\}$，那么当 $\widetilde{y}(x) \in \{i_1, \cdots, i_j\}$ 时，我们定义 g_f、w_μ、$\beta_\mu(x) = \widetilde{y}(x)$；当 $\widetilde{y}(x) \notin \{i_1, \cdots, i_j\}$ 时，则定义 g_f、w_μ、$\beta_\mu(x) = i_1$。事件 $\mathbb{A} := \{x \mid g_{f,W_\mu,\beta_\mu}(x) \neq \widetilde{y}(x)\} = \{x \mid g_{f,W_\mu,\beta_\mu}(x) \neq \widetilde{y}(x)\} \cap \bigcup_{i=1}^K \widetilde{\mathbb{M}}_i \subset \mathbb{X}$ 是事件 $\mathbb{D} := \bigcup_{i=1}^K \bigcup_{j\neq i} \{x \mid \|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x)) - \mu_j(f)\|_{\mathcal{H}_k}\} \cap \widetilde{\mathbb{M}}_i$ 的子集，因为

$$
\begin{aligned}
& x \in \mathbb{A} \\
\iff & \arg\min_{i\in[K]} \|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k} \neq \widetilde{y}(x) \quad \text{and} \quad x \in \bigcup_{i=1}^K \widetilde{\mathbb{M}}_i \quad (\text{def. of } g_{f,W_\mu,\beta_\mu} \text{ and } g_{1\text{-NN}}) \\
\iff & x \in \bigcup_{j\neq\widetilde{y}(x)} \{x' \mid \|h(f(x')) - \mu_{\widetilde{y}(x)}(h_f)\|_{\mathcal{H}_k} \geq \|h(f(x')) - \mu_j(f)\|_{\mathcal{H}_k}\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and} \quad x \in \bigcup_{i=1}^K \widetilde{\mathbb{M}}_i \\
\iff & x \in \bigcup_{j\neq\widetilde{y}(x)} \{x' \mid \|h(f(x')) - \mu_{\widetilde{y}(x)}(h_f)\|_{\mathcal{H}_k} \geq \|h(f(x')) - \mu_j(f)\|_{\mathcal{H}_k}\} \cap \widetilde{\mathbb{M}}_{\widetilde{y}(x)} \\
\implies & x \in \bigcup_{i=1}^K \bigcup_{j\neq i} \{x' \mid \|h(f(x')) - \mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x')) - \mu_j(f)\|_{\mathcal{H}_k}\} \cap \widetilde{\mathbb{M}}_i = \mathbb{D}.
\end{aligned}
$$

定义 $\mathbb{L}_{ij} := \{c(\mu_j(f) - \mu_i(f)) \mid c \in \mathbb{R}\} \subset \mathcal{H}_k$，对每个 $i, j \in [K], i \neq j$ 均成立。由于每个 $\mathbb{L}_{ij}$都是 $\mathcal{H}_k$的闭子空间，因此对每个 $z \in \mathcal{H}_k$，都存在某个 $z_1 \in \mathbb{L}_{ij}$ 和 $z_2 \in \mathbb{L}_{ij}^\perp$（其中 $\mathbb{L}_{ij}^\perp$ 是 $\mathbb{L}_{ij}$的正交补空间），使得 z 具有唯一分解$z = z_1 + z_2$。这里定义投影 $\widetilde{\pi}_{ij}\colon \mathcal{H}_k \to \mathbb{L}_{ij}$ 为 $\widetilde{\pi}_{ij}(z) = z_1$，并将平移投影 π_{ij} 定义为 $\pi_{ij}\colon \mathcal{H}_k \to \mathcal{H}_k, \pi_{ij}(z) := \widetilde{\pi}_{ij}(z - \mu_i(f)) + \mu_i(f)$。根据定义可得$\|\pi_{ij}(z) - \mu_i(f)\|_{\mathcal{H}_k} \leq \|z - \mu_i(f)\|_{\mathcal{H}_k}$，以及 $\|\pi_{ij}(z) - \mu_j(f)\|_{\mathcal{H}_k} \leq \|z - \mu_j(f)\|_{\mathcal{H}_k}$。

为方便起见，以下我们使用缩写 $\Delta_{ij} := \Delta_{ij}(f) = \|\mu_i(f) - \mu_j(f)\|_{\mathcal{H}_k}^2$。使用 $\pi_{ij}, i, j \in [K], i \neq j$，事件 $\mathbb{D}$ 可分解为，

$$
\begin{aligned}
\mathbb{D} = & \underbrace{\left(\mathbb{D} \cap \left(\bigcup_{i=1}^K \bigcup_{j\neq i} \left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right)\right)}_{=\mathbb{D}_1} \\
& \cup \underbrace{\left(\mathbb{D} \cap \left(\bigcup_{i=1}^K \bigcup_{j\neq i} \left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right)^c\right)}_{=\mathbb{D}_2}.
\end{aligned}
$$

对于 $\mathbb{D}_1$，我们有

$$
\begin{aligned}
& P_{\mathbb{X}}(\mathbb{D}_1) \\
& \leq P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\left(\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) && \text{(the union bound)} \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\left(\left\{x \mid -\|\pi_{ij}(h(f(x))) - \mu_i(f)\|_{\mathcal{H}_k} + \|\mu_i(f) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) \\
& && \text{(triangle ineq.)} \\
& = \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\left(\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_i(f)\|_{\mathcal{H}_k} \geq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right) \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} \frac{4}{\Delta_{ij}}\mathbb{E}\left[\|\pi_{ij}(h(f(x))) - \mu_i(f)\|_{\mathcal{H}_k}^2; \widetilde{\mathbb{M}}_i\right] && \text{(Markov's ineq.)} \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} \frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k}^2; \widetilde{\mathbb{M}}_i\right] && \text{(def. of } \pi_{ij}) \\
& \leq \sum_{i=1}^{K}\sum_{j\neq i} \frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x)) - \mu_i(f)\|_{\mathcal{H}_k}^2; \mathbb{M}_i\right]. && \text{(def. of } \widetilde{\mathbb{M}_i})
\end{aligned}
$$

对于 $\mathbb{D}_2$，我们注意到可以将其改写为:

$$
\begin{aligned}
& P_{\mathbb{X}}(\mathbb{D}_2) \\
& = P_{\mathbb{X}}\left(\mathbb{D} \cap \left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} \leq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i\right)^c\right) \\
& = P_{\mathbb{X}}\left(\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\{x \mid \|h(f(x)) - \mu_i(f)\|_2 \geq \|h(f(x)) - \mu_j(f)\|_2\} \cap \widetilde{\mathbb{M}}_i\right) \cap \right. \\
& \qquad \left.\left(\bigcap_{i=1}^{K}\bigcap_{j\neq i}\left\{x \mid \|\pi_{ij}(h(f(x))) - \mu_j(f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cup \widetilde{\mathbb{M}}_i^c\right)\right) \\
& = P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\bigcap_{i'=1}^{K}\bigcap_{j'\neq i'}\left(\{x \mid \|h(f(x)) - \mu_i(f)\|_2 \geq \|h(f(x)) - \mu_j(f)\|_2\} \cap \widetilde{\mathbb{M}}_i \cap \right.\right. \\
& \qquad \left.\left.\left(\left\{x \mid \|\pi_{i'j'}(h(f(x))) - \mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{i'j'}^{\frac{1}{2}}\right\} \cup \widetilde{\mathbb{M}}_{i'}^c\right)\right)\right)
\end{aligned}
$$

利用上述结果，我们得到

$$
\begin{aligned}
& P_{\mathbb{X}}(\mathbb{D}_2) \\
\leq\ & P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}(\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x))-\mu_j(f)\|_{\mathcal{H}_k} \right. \\
& \qquad \left. \text{and } \|\pi_{ij}(h(f(x)))-\mu_j(f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\} \cap \widetilde{\mathbb{M}}_i)\right) && (17) \\
\leq\ & P_{\mathbb{X}}\left(\bigcup_{i=1}^{K}\bigcup_{j\neq i}\left(\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k} \geq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\} \cap \widetilde{\mathbb{M}}_i\right)\right) && (\text{def. of } \pi_{ij}) \\
\leq\ & \sum_{i=1}^{K}\sum_{j\neq i} P_{\mathbb{X}}\left(\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k} \geq \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\} \cap \widetilde{\mathbb{M}}_i\right) && (\text{the union bound}) \\
\leq\ & \sum_{i=1}^{K}\sum_{j\neq i}\frac{4}{\Delta_{ij}}\mathbb{E}\left[\|\pi_{ij}(h(f(x)))-\mu_i(f)\|_{\mathcal{H}_k}^2; \widetilde{\mathbb{M}}_i\right] && (\text{Markov's ineq.}) \\
\leq\ & \sum_{i=1}^{K}\sum_{j\neq i}\frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}^2; \widetilde{\mathbb{M}}_i\right] && (\text{def. of } \pi_{ij}) \\
\leq\ & \sum_{i=1}^{K}\sum_{j\neq i}\frac{4}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}^2; \mathbb{M}_i\right]. && (\text{def. of } \widetilde{\mathbb{M}}_i)
\end{aligned}
$$

这里，我们证明（17）。首先，固定 i，$j \in [K]$，其中 $i \neq j$。对于 i'，$j' \in [K]$ 满足 $i' \neq j'$的情形，我们考虑以下两种情况。

- 如果 $i' = i$ 且 $j' = j$，那么 $\widetilde{\mathbb{M}}_i \cap \widetilde{\mathbb{M}}_i^c = \emptyset$，这意味着

$$
\begin{aligned}
& \{x \mid \|h(f(x))-\mu_i(f)\|_2 \geq \|h(f(x))-\mu_j(f)\|_2\} \cap \widetilde{\mathbb{M}}_i \cap \left(\left\{x \mid \|\pi_{i'j'}(h(f(x)))-\mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{i'j'}^{\frac{1}{2}}\right\} \cup \widetilde{\mathbb{M}}_{i'}^c\right) \\
& \quad = \left\{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x))-\mu_j(f)\|_{\mathcal{H}_k} \text{ and } \|\pi_{ij}(h(f(x)))-\mu_j(f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\right\} \cap \widetilde{\mathbb{M}}_i.
\end{aligned}
$$

- 如果 $i' \neq i$ 或 $j' \neq j$，那么

$$
\begin{aligned}
& \{x \mid \|h(f(x))-\mu_i(f)\|_2 \geq \|h(f(x))-\mu_j(f)\|_2\} \cap \widetilde{\mathbb{M}}_i \cap \left(\left\{x \mid \|\pi_{i'j'}(h(f(x)))-\mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{i'j'}^{\frac{1}{2}}\right\} \cup \widetilde{\mathbb{M}}_{i'}^c\right) \\
& \subset \mathbb{M}_i.
\end{aligned}
$$

因此，

$$\bigcap_{i'=1}^{K}\bigcap_{j'\neq i'}\Bigg(\{x \mid \|h(f(x))-\mu_i(f)\|_2 \geq \|h(f(x))-\mu_j(f)\|_2\} \\ \cap\widetilde{\mathbb{M}}_i \cap \left(\left\{x \mid \|\pi_{i'j'}(h(f(x)))-\mu_{j'}(h_f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{i'j'}^{\frac{1}{2}}\right\} \cup \widetilde{\mathbb{M}}_{i'}^{c}\right)\Bigg) \\ \subset \{x \mid \|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k} \geq \|h(f(x))-\mu_j(f)\|_{\mathcal{H}_k} \text{以及 } \|\pi_{ij}(h(f(x)))-\mu_j(f)\|_{\mathcal{H}_k} > \frac{1}{2}\Delta_{ij}^{\frac{1}{2}}\} \cap \widetilde{\mathbb{M}}_i.$$

综合所有结果，我们得到

$$\begin{aligned} P_{\mathbb{X}}(\mathbb{A}) &\leq P_{\mathbb{X}}(\mathbb{D}) \\ &\leq P_{\mathbb{X}}(\mathbb{D}_1)+P_{\mathbb{X}}(\mathbb{D}_2) \\ &\leq \sum_{i=1}^{K}\sum_{j\neq i}\frac{8}{\Delta_{ij}}\mathbb{E}\left[\|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}^2;\mathbb{M}_i\right] \\ &\leq \frac{8(K-1)}{\Delta_{\min}(f)}\sum_{i=1}^{K}\mathbb{E}[\|h(f(x))-\mu_i(f)\|_{\mathcal{H}_k}^2;\mathbb{M}_i] \\ &\leq \frac{8(K-1)}{\Delta_{\min}(f)}\sum_{i=1}^{K}\frac{1}{P_{\mathbb{X}}(\mathbb{M}_i)}\mathbb{E}_{x,x^-}[\|h(f(x))-h(f(x^-))\|_{\mathcal{H}_k}^2;\mathbb{M}_i\times\mathbb{M}_i] \quad \text{(Jensen's inequality)} \\ &\leq \frac{8(K-1)}{\Delta_{\min}(f)\cdot\min_{i\in[K]}P_{\mathbb{X}}(\mathbb{M}_i)}\mathfrak{a}(f), \end{aligned}$$

从而完成证明。 □

定理 *2* 的证明。 根据 y 的定义，可以保证集合 $\{x\in\mathbb{X}\mid y(x)=i\}$ 对于$i=1$、$\cdots$、K两两不相交，并满足关系 $\{x\in\mathbb{X}\mid y(x)$。 $=i\}\subseteq\mathbb{M}_i$因此，对于每个 $i\in[K]$，定理 5 均可应用于此情形，从而得到所需结果。 □

第 5.3 节中的证明

D.1 定理3的证明

首先，我们证明定理3。在此之前，我们先给出以下定理，它是定理3证明的一部分。

定理 6。 设 $(X_1,\ X_1')\ \cdots,\ (X_n,\ X_n')$ 为第 *5.3* 节中引入的随机变量。假设 *1* 成立，并假设 n 为偶数。则以至少 $1-\varepsilon$ 的概率，以下不等式成立：

$$\sup_{f\in\mathcal{F}}\left(-\frac{1}{n(n-1)}\sum_{i\neq j}k(f(X_i),f(X_j'))+\mathbb{E}_{X,X^-}\left[k(f(X),f(X^-))\right]\right)\leq 2\rho\mathfrak{R}_{n/2}^{-}(\mathcal{Q};s^*)+\sqrt{\frac{10b^2\log(1/\varepsilon)}{n}},$$

其中，我们将 $\mathfrak{R}^{-}_{n/2}(\mathcal{Q};\ s^*)$ 定义为次数为 n 的对称群 S_n：

$$\mathfrak{R}^{-}_{n/2}(\mathcal{Q};s^*) := \max_{s\in S_n} \mathbb{E}_{\substack{X,X' \\ \sigma_{1:(n/2)}}} \left[\sup_{f\in\mathcal{F}} \frac{2}{n} \sum_{i=1}^{n/2} \sigma_i f(X_{s(2i-1)})^\top f(X'_{s(2i)}) \right].$$

我们指出，与典型的泛化误差界（例如，Mohri 等人（2018）中的定理 3.3）相比，在定理 6 中，我们需要处理更为精细的技术问题，因为在我们的设定中，X_1、X'_1、$\cdots$、X_n、X'_n 彼此不一定独立。定理 6 的证明见附录 D.4。

现在，我们可以证明定理 3。

定理 *3* 的证明。 首先观察到，

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left(-\widehat{L}_{\mathrm{KCL}}(f;\lambda) + L_{\mathrm{KCL}}(f;\lambda) \right) \\
&= \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} k(f(X_i), f(X'_i)) - \frac{\lambda}{n(n-1)} \sum_{i\neq j} k(f(X_i), f(X'_j)) \right. \\
&\qquad\qquad \left. - \mathbb{E}_{X,X^+}\left[k(f(X), f(X^+))\right] + \lambda \mathbb{E}_{X,X^-}\left[k(f(X), f(X^-))\right] \right) \\
&\leq \underbrace{\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} k(f(X_i), f(X'_i)) - \mathbb{E}_{X,X^+}\left[k(f(X), f(X^+))\right] \right)}_{\text{(i)}} \\
&\qquad + \lambda \underbrace{\sup_{f\in\mathcal{F}} \left(-\frac{1}{n(n-1)} \sum_{i\neq j} k(f(X_i), f(X'_j)) + \mathbb{E}_{X,X^-}\left[k(f(X), f(X^-))\right] \right)}_{\text{(ii)}}.
\end{aligned}$$

我们定义函数空间 $\mathcal{K} := \{k(f(\cdot)\ f(\cdot)) : \mathbb{X}\times\mathbb{X}\to\mathbb{R} \mid f\in\mathcal{F}\}$。于是，$\mathcal{K}$一致有界，其常数为 $b = \sup_{z,z\in\mathbb{S}^{d-1}} |k(z,z')|$。这里注意到，$b < +\infty$ 成立，这是因为 k 连续且 $\mathbb{S}^{d-1}$ 紧致；参见第 2.1 节。根据一致大数定律（Mohri 等人（2018）的定理 3.3），以至少 $1-\varepsilon/2$ 的概率，我们有

$$\text{(i)} \leq 2\mathfrak{R}^{+}_{n}(\mathcal{K}) + \sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}}.$$

由于 k 由 $k(x,\ x') = \psi(x^\top x')$ 表示，其中 ρ是一个 -Lipschitz 函数 ψ，由假设 1，应用 Talagrand 引理（Shalev-Shwartz 和 Ben-David (2014年) 中的引理 26.9）可得$\mathfrak{R}^{+}_{n}(\mathcal{K}) \leq \rho\mathfrak{R}^{+}_{n}(\mathcal{Q})$。因此，以至少 $1-\varepsilon/2$ 的概率，有

$$\text{(i)} \leq 2\rho\mathfrak{R}^{+}_{n}(\mathcal{Q}) + \sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}}.$$

对于（ii），由定理 6，以至少 $1-\varepsilon/2$ 的概率，有

$$\text{(ii)} \leq 2\rho\mathfrak{R}^{-}_{n/2}(\mathcal{Q}; s^*) + \sqrt{\frac{10b^2 \log(2/\varepsilon)}{n}}.$$

因此，以至少 $1-\varepsilon$ 的概率，我们有：

$$\sup_{f\in\mathcal{F}} \left(-\widehat{L}_{\mathrm{KCL}}(f;\lambda) + L_{\mathrm{KCL}}(f;\lambda)\right) \leq 2\rho\mathfrak{R}_n(\mathcal{Q}) + \sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}} + \lambda\sqrt{\frac{10b^2 \log(2/\varepsilon)}{n}}, \tag{18}$$

其中 $\mathfrak{R}_n(\mathcal{Q}) := \mathfrak{R}^{+}_n(\mathcal{Q}) + \lambda\mathfrak{R}^{-}_{n/2}(\mathcal{Q}; s^*)$。

注意，与上述概率界的证明相同，我们有以下不等式：以至少 $1-\varepsilon$ 的概率，

$$\sup_{f\in\mathcal{F}} \left(\widehat{L}_{\mathrm{KCL}}(f;\lambda) - L_{\mathrm{KCL}}(f;\lambda)\right) \leq 2\rho\mathfrak{R}_n(\mathcal{Q}) + \sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}} + \lambda\sqrt{\frac{10b^2 \log(2/\varepsilon)}{n}}. \tag{19}$$

因此，令 $\widehat{f}$ 为 $L_{\mathrm{KCL}}(f;\lambda)$ 的极小化器，则由（18）和（19），以至少$1-2\varepsilon$ 的概率，我们有

$$L_{\mathrm{KCL}}(\widehat{f};\lambda) \leq L_{\mathrm{KCL}}(f;\lambda) + 4\rho\mathfrak{R}_n(\mathcal{Q}) + 2\sqrt{\frac{2b^2 \log(2/\varepsilon)}{n}} + 2\lambda\sqrt{\frac{10b^2 \log(2/\varepsilon)}{n}},$$

其中，根据 $\widehat{f}$ 的定义，我们注意到 $L_{\mathrm{KCL}}(\widehat{f};\lambda) \leq L_{\mathrm{KCL}}(f;\lambda)$。因此，证明完成。 □

D.2 拉德马赫复杂度的上界

在本节中，为简便起见，我们考虑如下情形：对于每个 $f\in\mathcal{F}$，存在唯一函数 $f_0\in\mathcal{F}_0$，使得 $f(x) = f_0(x)/\|f_0(x)\|_2$ 对每个 $x\in\mathbb{X}$均成立。首先回顾次高斯过程的定义：

定义 4（引自Wainwright （2019年）中的定义 5.16）。零均值随机变量的集合 $\{X_\theta, \theta\in\mathbb{T}\}$ 是关于度量 ρ_X 在 $\mathbb{T}$ 上的次高斯过程，当且仅当

$$\mathbb{E}\left[e^{\lambda(X_\theta - X_{\widetilde{\theta}})}\right] \leq e^{\frac{\lambda^2\rho_X^2(\theta,\widetilde{\theta})}{2}} \quad \text{for all } \theta,\widetilde{\theta}\in\mathbb{T}, \text{ and } \lambda\in\mathbb{R}.$$

接下来，我们通过链式方法对拉德马赫复杂度进行上界估计（Wainwright （2019年）中的定理 5.22）。

命题 2。假设 n 为偶数。对于 $\mathfrak{R}_n(\mathcal{Q})$，我们有如下上界：

$$\mathfrak{R}_n(\mathcal{Q}) \leq \frac{64(1+\sqrt{2}\lambda)}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}} \int_0^{Cd} \sqrt{\log \mathfrak{C}(u;\mathcal{F}_0, \|\cdot\|_\infty)}\, du,$$

其中，$\|f_0\|_\infty := \sup_{x\in\mathbb{X}} \|f_0(x)\|_2$ *for* $f_0\in\mathcal{F}_0$，$\mathfrak{m}(\mathcal{F}_0)$ 在第 *2.1* 节中定义，C 是一个与 d, n, λ 独立的常数，而 $\mathfrak{C}(u;\mathcal{F}_0, \|\cdot\|_\infty)$ 是 $(\mathcal{F}_0, \|\cdot\|_\infty)$ 的 u-覆盖数（关于覆盖数的定义，参见 *Wainwright*（*2019*）中的定义 *5.1*）。

命题 2 的证明。 在此证明中，我们沿用 Tu 等人 (2019年) 的证明思路（参见 Tu 等人 (2019年) 中的引理 5）。由于我们的设定不同于Tu 等人 (2019年，我们需要修改证明并加入几种新技巧。定义

$$Z_{f_0} := \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \sum_{i=1}^{n} \sigma_i q(f_0(X_i), f_0(X_i')),$$

其中 σ_1， $\cdots$， σ_n 是彼此独立且与每个（X_i， X_i'）相互独立的拉德马赫随机变量， $i \in [n]$， $f_0 \in \mathcal{F}_0$， （X_1， X_1） $\cdots$， （X_n， X_n'）是第 5.3 节中定义的随机向量，并且

$$q(z, z') = \frac{z^\top z'}{\|z\|_2 \cdot \|z'\|_2} \quad z, z' \in \mathbb{S}^{d-1}.$$

此外，回顾一下第 2.1 节中引入的关于 $\mathcal{F}_0$ 的假设：对于每个 $f \in \mathcal{F}$，存在唯一的函数 $f_0 \in \mathcal{F}_0$ ，使得对于每个 $x \in \mathbb{X}$， $f(x) = f_0(x)/\|f_0(x)\|_2$ 。我们证明$\{Z_{f_0}\}_{f_0 \in \mathcal{F}_0}$ 是一个次高斯过程如下：注意，对于每个 $f_{1,0}$， $f_{2,0} \in \mathcal{F}_0$，

$$\begin{aligned}
&\frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left|\sigma_i(q(f_{1,0}(X_i), f_{1,0}(X_i')) - q(f_{2,0}(X_i), f_{2,0}(X_i')))\right| \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left|f_1(X_i)^\top f_1(X_i') - f_2(X_i)^\top f_2(X_i')\right| && \text{(def. of } \sigma_i\text{)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left(\left|f_1(X_i)^\top f_1(X_i') - f_1(X_i)^\top f_2(X_i')\right| + \left|f_1(X_i)^\top f_2(X_i') - f_2(X_i)^\top f_2(X_i')\right|\right) \\
& && \text{(triangle ineq.)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left(\|f_1(X_i)\|_2 \|f_1(X_i') - f_2(X_i')\|_2 + \|f_1(X_i) - f_2(X_i)\|_2 \|f_2(X_i')\|_2\right) \\
& && \text{(Cauchy-Schwarz ineq.)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{2\sqrt{n}} \left(\|f_1(X_i') - f_2(X_i')\|_2 + \|f_1(X_i) - f_2(X_i)\|_2\right) && \text{(def. of } f_1, f_2\text{)} \\
&\leq \frac{\mathfrak{m}(\mathcal{F}_0)}{\sqrt{n}} \sup_{x \in \mathbb{X}} \|f_1(x) - f_2(x)\|_2 \\
&= \frac{\mathfrak{m}(\mathcal{F}_0)}{\sqrt{n}} \sup_{x \in \mathbb{X}} \left\| \frac{f_{1,0}(x)}{\|f_{1,0}(x)\|_2} - \frac{f_{2,0}(x)}{\|f_{2,0}(x)\|_2} \right\|_2 && \text{(the uniqueness of } f_{1,0}, f_{2,0}\text{)} \\
&\leq \frac{1}{\sqrt{n}} \|f_{1,0} - f_{2,0}\|_\infty. && \text{(def. of } \mathfrak{m}(\mathcal{F}_0)\text{)}
\end{aligned}$$

因此，我们有

$$\mathbb{E}_{X_{1:n}, X'_{1:n}, \sigma_{1:n}} \left[\exp\left(t(Z_{f_{1,0}} - Z_{f_{2,0}})\right)\right] \leq \exp\left(\frac{t^2}{2n} \|f_{1,0} - f_{2,0}\|_\infty^2\right)^n = \exp\left(\frac{t^2}{2} \|f_{1,0} - f_{2,0}\|_\infty^2\right).$$

这表明 $\{Z_{f_0}\}_{f \in \mathcal{F}_0}$ 是一个范数为 $\|\cdot\|_\infty$的次高斯过程。这里注意，$\sup_{f_{1,0}, f_{2,0} \in \mathcal{F}_0} \|f_{1,0} - f_{2,0}\|_\infty \leq C\sqrt{d}$ 对于某个常数 $C \in \mathbb{R}$，该常数与 d无关，因为 $\mathcal{F}_0$ 是一致有界的。通过使用 Wainwright（2019年）中的链式定理（定理 5.22），我们得到

$$\mathfrak{R}_n^+(\mathcal{Q}) \leq \frac{64}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}} \int_0^{C\sqrt{d}} \sqrt{\log \mathfrak{C}(u; \mathcal{F}_0, \|\cdot\|_\infty)} du.$$

对于 $\mathfrak{R}^{-}_{n/2}(\mathcal{Q})$，以类似的方式可得，

$$\mathfrak{R}^{-}_{n/2}(\mathcal{Q}) \leq \frac{64\sqrt{2}}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}} \int_0^{C\sqrt{d}} \sqrt{\log \mathfrak{C}(u; \mathcal{F}_0, \|\cdot\|_\infty)} du.$$

因此，我们有

$$\mathfrak{R}_n(\mathcal{Q}) \leq \frac{64(1+\sqrt{2}\lambda)}{\mathfrak{m}(\mathcal{F}_0)\sqrt{n}} \int_0^{C\sqrt{d}} \sqrt{\log \mathfrak{C}(u; \mathcal{F}_0, \|\cdot\|_\infty)} du,$$

并完成证明。 □

上述上界中的积分通常称为达德利熵积分（Wainwright，2019年）。命题2通过链式方法使推导泛化界变得更加容易，因为只需计算函数空间 $\mathcal{F}_0$ 的达德利熵积分，而无需计算评论家函数空间 $\mathcal{Q}$。

这里，将 $\mathfrak{D}(\mathcal{F}_0, \|\cdot\|_\infty)$ 记为关于 $(\mathcal{F}_0, \|\cdot\|_\infty)$ 的达德利熵积分，即

$$\mathfrak{D}(\mathcal{F}_0, \|\cdot\|_\infty) = \int_0^{C\sqrt{d}} \sqrt{\log \mathfrak{C}(u; \mathcal{F}_0, \|\cdot\|_\infty)} du.$$

Tu 等人（2019年）证明，如果 $\mathcal{F}_0$ 是前馈（深度）神经网络的函数空间，其中每个神经网络的权重矩阵的范数均由某个通用常数界定，且使用在原点处为零的Lipschitz激活函数，那么 $\mathfrak{D}(\mathcal{F}_0, \|\cdot\|_\infty) < +\infty$成立。基于这一事实，我们引入：

假设 3. 达德利熵积分 $\mathfrak{D}(\mathcal{F}_0, \|\cdot\|_\infty)$ 是有限的，并且 $\mathfrak{R}_n(\mathcal{Q}) \leq O((1+\lambda)/\sqrt{n})$成立。 –

因此，我们得到泛化误差界。

推论2. 假设 *1*、*3* 成立，且 n 为偶数。那么，以至少$1-\varepsilon$ 的概率，其中 $\varepsilon > 0$，我们有

$$L_{\mathrm{KCL}}(f;\lambda) \leq \widehat{L}_{\mathrm{KCL}}(f;\lambda) + O\left(\frac{(1+\lambda)\left(1+\sqrt{\log(2/\varepsilon)}\right)}{\sqrt{n}}\right).$$

证明. 根据定理3 和假设3。 □

D.3 相关随机变量的 McDiarmid 不等式有用结果

在证明定理 6之前，我们需要准备若干定义和一个已有结果。以下三个定义引自 Zhang等人(2019年)。

定义5 （依赖图，摘自 Zhang 等人 （2019年））。 无向图 G 称为随机向量 $\mathbf{X} = (X_1, \cdots, X_n)$，如果

1. V（G）$=[n]$*2.* 如果 I、$J\subset[n]$ 在 G中不相邻，那么 $\{X_i\}_{i\in I}$ 和 $\{X_j\}_{j\in J}$ 相互独立。

定义 *6*（森林近似，引自 *Zhang* 等人（*2019*）中的定义 *3.4*）。给定一个图 G、一个森林 F 和一个映射 ϕ*:* $V(G)\to V(F)$，若对于任意 $\langle u,v\rangle\in E(G)$，$\phi(u)=\phi(v)$ 或 $\langle\phi(u)\ \phi(v)\rangle\in E(F)$，则称 (ϕ,F) 是 G 的一个森林近似。令 $\Phi(G)$ 表示 G 的森林近似集合。

定义 7（森林复杂度，引自Zhang 等人（2019年））。给定一个图 G 以及任意森林近似（ϕ，F）$\in\Phi$（G），其中 F由树 $\{T_i\}_{i\in[k]}$组成，令

$$\lambda_{(\phi,F)}=\sum_{\langle u,v\rangle\in E(F)}\left(|\phi^{-1}(u)|+|\phi^{-1}(v)|\right)^2+\sum_{i=1}^{k}\min_{u\in V(T_i)}|\phi^{-1}(u)|^2.$$

我们称

$$\Lambda(G)=\min_{(\phi,F)\in\Phi(G)}\lambda_{(\phi,F)}$$

为图 G 的森林复杂度。

Zhang 等人（2019年）证明了以下结果，这是 McDiarmid 不等式（McDiarmid，1989）针对相关随机变量的推广。

定理 7 (引自 Zhang 等人 (2019年))。 假设 $f:\mathbf{\Omega}\to\mathbb{R}$ 是一个**c-***Lipschitz* 函数，且 G 是随机向量 $\mathbf{X}$（取值于 $\mathbf{\Omega}$）的依赖图。对于任意 $t>0$，以下不等式成立：

$$\Pr(f(\mathbf{X})-\mathbf{E}[f(\mathbf{X})]\geq t)\leq\exp\left(-\frac{2t^2}{\Lambda(G)\|\mathbf{c}\|_\infty^2}\right).$$

注意，在上述定理中，$f:\mathbf{\Omega}\to\mathbb{R}$ 称为<c>-Lipschitz，如果 $|f(\mathbf{x})-f(\mathbf{x}')|\leq\sum_{i=1}^{p}\mathbf{c}_i\mathbb{1}_{\{\mathbf{x}_i\neq\mathbf{x}'_i\}}$ 对每个$\mathbf{x},\mathbf{x}'\in\mathbf{\Omega}$均成立，其中$\mathbf{\Omega}\subset\mathbb{R}^p$ 对某个 $p\in\mathbb{N}$成立。

D.4 定理证明6

我们利用附录 6 D.3 中的内容D.3来证明定理。回顾第 5.3节中引入的随机变量定义：$(X_1,X'_1)\cdots,(X_n,X'_n)$ 是根据密度为 $w(x,x')$ 的联合概率分布独立采样得到的随机变量对，其中，对于每一对不同的索引 X_i 和X'_j，i，$j\in[K]$都是独立的。根据定义，以下结论成立。

引理 3.设 G_n 为由随机向量$(X_1,X'_1,\cdots,X_n,X'_n)$定义的依赖图，其中 G_n 中的边定义如下：对于任意 i、$j\in[n]$、X_i和X_j，它们不相连；X_i 和 X'_j 之间有边，当且仅当 $i=j$。于是，我们有 $\Lambda(G_n)\leq 5n$。

证明。设 ϕ: $G_n \to G_n$ 为恒等映射。根据定义， G_n 可以分解为树 $\{T_i\}_{i\in[n]}$，其中对于每个 $i \in [n]$, $V(T_i) = \{X_i, X_i'\}$。设 F为由树$\{T_i\}_{i\in[n]}$组成的森林。于是，我们有 $\lambda_{(\phi,F)}= 5n$，这意味着 $\Lambda(G_n)$ $\leq \lambda_{(\phi,F)} \leq 5n$。 □

定理 *6*的证明。本证明旨在以高概率给出以下量的上界:

$$\sup_{f\in\mathcal{F}}\left(-\frac{1}{n(n-1)}\sum_{i\neq j}k(f(X_i),f(X_j'))+\mathbb{E}_{X,X^-}\left[k(f(X_i),f(X^-))\right]\right).$$

然而，正如前文所述，标准论证（例如，参见 Mohri 等人（2018）中的定理 3.3）不适用于这种情况，因为根据我们的问题设定， $k(f(X_i)\ f(X_j'))$, $i, j \in [n]$, $i \neq j$彼此之间不一定独立。为避免这一问题，我们采用 Zhang 等人（2019）提出的适用于相关随机变量的 McDiarmid 不等式。以下证明主要基于 Mohri 等人（2018）中的定理 3.3，但根据 Zhang 等人（2019）结果的应用作了相应修改。设 $X_1, X_1', \cdots, X_n, X_n'$ 是与原随机变量 $X_1, X_1'. \cdots, X_n, X_n'$独立同分布的随机变量。定义可测函数$F(f) := F(f)(x_1, x_1', \cdots, x_n, x_n')$，其定义域为 $\mathbb{X}^{2n}$，并令其为

$$F(f) := \frac{1}{n(n-1)}\sum_{i\neq j}k(f(x_i),f(x_j'))-\mathbb{E}_{X,X^-}\left[k(f(X),f(X^-))\right].$$

为简便起见，记

$$F(f)_{x_\ell} := \frac{1}{n(n-1)}\left(\sum_{j:j\neq\ell}k(f(\widetilde{x}_\ell),f(x_j'))+\sum_{\substack{i,j:i\neq j\\ i\neq\ell}}k(f(x_i),f(x_j'))\right)-\mathbb{E}_{X,X^-}\left[k(f(X),f(X^-))\right].$$

类似地 ，我们也使用记号 $F(f)^{x_\ell'}$。令 $j \in [n]$。那么，对于每个 $f \in \mathcal{F}$，我们有

$$\begin{aligned}F(f)-\sup_{f\in\mathcal{F}}F(f)_{X_j}\leq F(f)-F(f)_{X_j}&\leq\left|F(f)-F(f)_{X_j}\right|\\&\leq\left|\frac{1}{n(n-1)}\sum_{i\in[n],i\neq j}\left(k(f(X_j),f(X_i'))-k(f(\widetilde{X}_j),f(X_i'))\right)\right|\\&\leq\frac{1}{n(n-1)}\cdot 2(n-1)b=\frac{2b}{n},\end{aligned}$$

其中 $b := \sup_{z,z'\in\mathbb{S}^{d-1}}|k(z,z')|$。因此，$\sup_{f\in\mathcal{F}}F(f)\ -\sup_{f\in\mathcal{F}}F(f)_{X_j}\ \leq\ \frac{2b}{n}$。重复应用相同的论证数次后，$\sup_{f\in\mathcal{F}}F(f)$ 满足定理 7 的假设。因此，根据定理 7（即 Zhang 等人（2019）中的定理 3.6）和引理 3，以至少 $1-\varepsilon$ 的概率，我们有

$$\sup_{f\in\mathcal{F}}F(f)\leq\mathbb{E}\left[\sup_{f\in\mathcal{F}}F(f)\right]+\sqrt{\frac{10b^2\log(1/\varepsilon)}{n}}.\tag{20}$$

设 $\sigma_{1:n/2} := (\sigma_1, \cdots, \sigma_{n/2})$ 是一个随机向量，其中每个分量都是拉德马赫随机变量~ ~（即以各为 $1/2$ 的概率取值 ± 1 的随机变量），并设 $X_{1:n}, X'_{1:n}$ 分别是随机向量 $X_{1:n}, X'_{1:n}$ 的独立同分布副本。记 $m = n/2 \in \mathbb{N}$。于是，

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}} F(f)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n(n-1)}\sum_{i\neq j} k(f(X_i), f(X'_j)) - \mathbb{E}_{X,X^-}\left[k(f(X), f(X^-))\right]\right)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n(n-1)}\sum_{i\neq j} k(f(X_i), f(X'_j)) - \mathbb{E}_{\widetilde{X}_{1:n},\widetilde{X}'_{1:n}}\left[\frac{1}{n(n-1)}\sum_{i\neq j} k(f(\widetilde{X}_i), f(\widetilde{X}'_j))\right]\right)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n!m}\sum_{s\in S_n}\left(\sum_{i=1}^{m} k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - \mathbb{E}_{\widetilde{X}'_{1:n},\widetilde{X}'_{1:n}}\left[\sum_{i=1}^{m} k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right]\right)\right)\right] \qquad (21) \\
&\leq \frac{1}{n!}\sum_{s\in S_n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - \mathbb{E}_{\widetilde{X}_{1:n},\widetilde{X}'_{1:n}}\left[\frac{1}{m}\sum_{i=1}^{m} k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right]\right)\right] \\
&\leq \frac{1}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \widetilde{X}_{1:n},\widetilde{X}'_{1:n}}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right)\right)\right] \\
&= \frac{1}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \widetilde{X}_{1:n},\widetilde{X}'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(k(f(X_{s(2i-1)}), f(X'_{s(2i)})) - k(f(\widetilde{X}_{s(2i-1)}), f(\widetilde{X}'_{s(2i)}))\right)\right)\right] \qquad (22) \\
&\leq \frac{2}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i k(f(X_{s(2i-1)}), f(X'_{s(2i)}))\right] \\
&\leq \frac{2\rho}{n!}\sum_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(X_{s(2i-1)})^\top f(X'_{s(2i)})\right] \qquad (23) \\
&\leq 2\rho\max_{s\in S_n}\mathbb{E}_{\substack{X_{1:n},X'_{1:n}\\ \sigma_{1:m}}}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(X_{s(2i-1)})^\top f(X'_{s(2i)})\right] \\
&= 2\rho\mathfrak{R}_m^-(\mathcal{Q}; s^*),
\end{aligned}
$$

其中，在 (21) 中，我们将 S_n 定义为 n 次对称群（关于其与 U 统计学中“独立同分布块求和”技术之平均值的关系，参见注 3；该技术由 Cl´emen¸con 等人~ ~（2008年）解释）。此外，在 (22) 中，对于每个 $s \in S_n$，随机向量 $(X_{s(2i-1)}, X_{s(2i)})$、$(X_{s(2i-1)}, X_{s(2i)})$（其中 $i = 1, \cdots, m$）独立同分布，因此可以应用标准对称化论证（Wainwright（2019年）的定理 4.10）。最后，在 (23) 中，在假设 1 下，我们应用 Talagrand 引理（Shalev-Shwartz 和 Ben-David 的引理 26.9

（2014年））。因此，我们以至少 $1-\varepsilon$ 的概率获得，

$$\sup_{f\in\mathcal{F}} F(f) \leq 2\rho\mathfrak{R}^{-}_{n/2}(\mathcal{Q};s^*) + \sqrt{\frac{10b^2\log(1/\varepsilon)}{n}}.$$

由此，我们得到了这一结论。 □

注 3. 在 (21) 定理 *6*的证明中，我们使用了以下恒等式：

$$\frac{1}{n(n-1)}\sum_{i\neq j} k(f(X_i),f(X'_j)) = \frac{1}{n!m}\sum_{s\in S_n}\sum_{i=1}^{m} k(f(X_{s(2i-1)},f(X'_{s(2i)})).$$

我们注意到，上述恒等式与 *Cl´emen¸con* 等人（*2008*年）在附录*A*中介绍的"独立同分布块求和"平均技术密切相关。与 *Cl´emen¸con* 等人（*2008*年）提出的技术一样，在本文的 *(21)* 式中，我们也将总和 $\sum_{i\neq j} k(f(X_i)\ f(X'_j))$ 分解为独立同分布随机变量之和。不过，我们要指出的是，总和$\sum_{i\neq j} k(f(X_i)\ f(X'_j))$ 的定义不同于 *Cl´emen¸con* 等人（*2008*年）提出的定义：事实上，在我们的情形下，随机变量 $f(X_1)\ f(X'_1)\ \cdots, f(X_n)\ f(X'_n)$ 不一定相互独立。为解决这一问题，我们将 *(21)* 式中的总和分解如下：对于 $2n$ 随机变量$X_1, X'_1, \cdots, X_n, X'_n$，我们构造元组 $(X_{s(1)}, X'_{s(2)}, \cdots, X_{s(n-1)}, X'_{s(n)})$，其中 $s\in S_n$，然后将所有分量 $\{\sum_{i=1}^{n} k(f(X_{s(2i-1)})\ f(X'_{s(2i)}))\}_{s\in S_n}$ 相加。

E 第5.4节的证明

定理的证明 $\widehat{4}$：首先应用定理*2* 于经验损失最小化器 f，我们有 e

$$L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y) \leq \frac{8(K-1)}{\Delta_{\min}(\widehat{f})\cdot\min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)}\mathfrak{a}(\widehat{f}). \tag{24}$$

根据定理1，我们有以下不等式：

$$\mathfrak{a}(\widehat{f}) \leq L_{\mathrm{KCL}}(\widehat{f};\lambda) + (1-\frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda\mathfrak{c}(\widehat{f}) + R(\lambda). \tag{25}$$

结合式（24）和（25），可得

$$L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y) \leq \frac{8(K-1)}{\Delta_{\min}(\widehat{f})\cdot\min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)}\left(L_{\mathrm{KCL}}(\widehat{f};\lambda) + (1-\frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda\mathfrak{c}(\widehat{f}) + R(\lambda)\right). \tag{26}$$

这里采用对最优分类损失或误差给出上界的标准技术（Arora $\widehat{}$等人，2019年；Ash 等人，2022年），分类误差 $L_{\mathrm{Err}}(f,\ W_\mu,\ \beta_\mu;\ y$的下界为

$$L_{\mathrm{Err}}(\widehat{f},W^*\beta^*;y) = \inf_{W,\beta} L_{\mathrm{Err}}(\widehat{f},W,\beta;y) \leq L_{\mathrm{Err}}(\widehat{f},W_\mu,\beta_\mu;y). \tag{27}$$

由（26）和（27）可得，

$$L_{\mathrm{Err}}(\widehat{f}, W^*, \beta^*; y) \leq \frac{8(K-1)}{\Delta_{\min}(\widehat{f}) \cdot \min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)} \left(L_{\mathrm{KCL}}(\widehat{f}; \lambda) + (1 - \frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda \mathfrak{c}(\widehat{f}) + R(\lambda) \right). \tag{28}$$

应用定理 3 于 (28)，我们得到：以至少 $1-2\varepsilon$ 的概率，

$$L_{\mathrm{Err}}(\widehat{f}, W_\mu, \beta_\mu; y) \lesssim L_{\mathrm{KCL}}(f; \lambda) + (1 - \frac{\delta}{2})\mathfrak{a}(\widehat{f}) - \lambda \mathfrak{c}(\widehat{f}) + R(\lambda) + 2\mathrm{Gen}(n, \lambda, \varepsilon),$$

其中 $\lesssim$ 省略了系数 $8(K-1)\Delta_{\min}(\widehat{f}) \cdot \min_{i\in[K]} P_{\mathbb{X}}(\mathbb{M}_i)$。因此，我们得到该结果。 □

F 附加信息、结果与讨论

F.1 满足假设 2

F.1.1 示例 1 的证明

我们将示例 1 中出现的若干断言表述为一个命题。

命题 3。设 $r > 0$、$K \in \mathbb{N}$ 和 v_1、$\cdots$、$v_K \in \mathbb{R}^p$。对于每个 $i \in [K]$，令 $\mathbb{B}_i \subset \mathbb{R}^p$ 为以点 v_i 为中心、半径为 r 的开球。假设 $\mathbb{B}_1$、$\cdots$、$\mathbb{B}_K$ 两两不相交。定义 $\mathbb{X} = \bigcup_{i=1}^K \mathbb{B}_i$、$\mathbb{X} = \mathbb{X}$，以及条件概率 $a(x|x) = \mathrm{vol}(\mathbb{B}_1)^{-1} \sum_{i=1}^K \mathbb{1}_{\mathbb{B}_i \times \mathbb{B}_i}(x,x)$，其中 $\mathrm{vol}(\mathbb{B}_1)$ 是 $\mathbb{R}^p$ 中 $\mathbb{B}_i$ 的体积。令 $p_{\mathbb{X}}(x) := (K\mathrm{vol}(\mathbb{B}_1))^{-1}$ 为 $P_{\mathbb{X}}$ 的概率密度函数。定义 $y : \mathbb{X} \to [K]$ 为 $y(x) = i$，若 $x \in \mathbb{B}_i$。那么，我们有以下性质：

1. $w(x) > 0$ 对每个 $x \in \mathbb{X}$ 均成立。2. 相似度 $(x, x'; \lambda) = K\mathbb{1}_{\bigcup_{i\in[K]} \mathbb{B}_i \times \mathbb{B}_i}(x, x') - \lambda$ 对每个 x、$x' \in \mathbb{X}$ 均成立。3

。令 $\delta \in (-\lambda, K-\lambda]$。那么，$\delta$，$K$，$\mathbb{B}_1$，$\cdots$，$\mathbb{B}_K$，且 y 满足假设 2。

证明。我们首先证明结论 1。根据 $w(x)$ 的定义，对于每个 $x \in \mathbb{B}_1$，我们有

$$w(x) = \int_{\mathbb{X}} a(x|\overline{x}) p_{\mathbb{X}}(\overline{x}) d\overline{x} = \int_{\mathbb{B}_1} \frac{1}{K\left(\mathrm{vol}(\mathbb{B}_1)\right)^2} d\overline{x} = \frac{1}{K\mathrm{vol}(\mathbb{B}_1)}.$$

类似地，对于每个 $i \in [K]$，我们得到 $w(x) = (K\text{体积}(\mathbb{B}_1))^{-1}$，对每个 $x \in \mathbb{B}_i$ 均成立。由于 $\mathbb{X} = \mathbb{X} = \bigcup_{i=1}^K \mathbb{B}_i$，我们有 $w(x) > 0$，对每个 $x \in \mathbb{X}$ 均成立。

接下来，我们证明命题 2。根据命题 1，函数 $\mathrm{sim}(x, x'; \lambda)$ 定义良好。要计算 $\mathrm{sim}(x, x'; \lambda)$，我们需要知道函数 $w(x, x')$。$w(x, x')$ 的计算过程

如下：

$$
\begin{aligned}
w(x,x') &= \int_{\overline{\mathbb{X}}} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\
&= \begin{cases} \int_{\mathbb{B}_i} \frac{1}{K(\mathrm{vol}(\mathbb{B}_1))^3} d\overline{x} & \text{if } x, x' \in \mathbb{B}_i \text{ for some } i \in [K] \\ 0 & \text{if } x \in \mathbb{B}_i \text{ and } x' \in \mathbb{B}_j \text{ for some } i \neq j \end{cases} \\
&= \begin{cases} \frac{1}{K(\mathrm{vol}(\mathbb{B}_1))^2} & \text{if } x, x' \in \mathbb{B}_i \text{ for some } i \in [K] \\ 0 & \text{if } x \in \mathbb{B}_i \text{ and } x' \in \mathbb{B}_j \text{ for some } i \neq j. \end{cases}
\end{aligned}
$$

因此，显然命题 2 成立。

最后，让我们证明命题 3。然而，由命题 2 可知，$\mathsf{sim}(x, x'; \lambda) \geq \delta$ 当且仅当 $x, x' \in \mathbb{B}_i$ 对某个 $i \in [K]$成立。此外，y 是良定义的，并且对于每个 $i \in [K]$，集合 $\{x \in \mathbb{X} \mid y(x) = i\} = \mathbb{B}_i$都是可测的。因此，命题 3 也成立，证明完毕。 □

F.1.2 簇重叠时的示例

这里，我们也讨论一个 $\mathbb{X}$ 中的簇存在部分重叠的示例。在以下命题中，为简单起见，我们考虑 $\mathbb{X}$中有两个簇的情形。

命题 *4.* 设 $r > 0$，且 $v_1, v_2 \in \mathbb{R}^p$。对于每个 $i \in \{1, 2\}$，令 $\mathbb{B}(v_i; r) \subset \mathbb{R}^p$ 为以点 v_i为中心、半径为 r 的开球。假设 $\|v_1 - v_2\|_2 = 3r$。定义 $\mathbb{X} = \mathbb{B}(v_1; r) \cup \mathbb{B}(v_2; r)$，$\mathbb{X} = \mathbb{B}(v_1; 2r) \cup \mathbb{B}(v_2; \overline{2r})$，以及 $a(x|x) = \mathrm{vol}(\mathbb{B}(v_1; 2r))^{-1} \sum_{i=1}^{2} \mathbb{1}_{\mathbb{B}(v_i;2r)\times\mathbb{B}(v_i;r)}(x, \overline{x})$。令 $p_{\mathbb{X}}(x) := (2 \cdot \mathrm{vol}(\mathbb{B}(v_1; r)))^{-1}$ 为 $P_{\mathbb{X}}$的概率密度函数。定义$\overline{\ }y: \mathbb{X} \to \{1, 2\}$ 为 $y(x) = 1$，若$x \in \mathbb{B}(v_1; 2r)$，且 $y(x) = 2$，若 $x \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$。那么，我们得到以下结果：

1. $w(x) > 0$ 对每个 $x \in \mathbb{X}$。
2. $\mathsf{sim}(x, x'; \lambda) = 2-\lambda$ 如果 $x, x' \in \mathbb{B}(v_1; 2r)\setminus\mathbb{B}(v_2; 2r)$ 或 $x, x' \in \mathbb{B}(v_2; 2r)\setminus\mathbb{B}(v_1; 2r)$，则 $\mathsf{sim}(x, x'; \lambda) = -\lambda$ 如果 $(x, x') \in (\mathbb{B}(v_1; 2r)\setminus\mathbb{B}(v_2; 2r))\times(\mathbb{B}(v_2; 2r)\setminus\mathbb{B}(v_1; 2r))$ 或 $(x, x') \in (\mathbb{B}(v_2; 2r)\setminus\mathbb{B}(v_1; 2r))\times(\mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r))$，并且 $\mathsf{sim}(x, x'; \lambda) = 1 - \lambda$ 否则。
3. 设 $\delta \in (-\lambda, 1 - \lambda]$。那么，$\delta$, K, $\mathbb{B}_1, \cdots, \mathbb{B}_K$，且 y 满足假设 2。

证明。设 $\mathbb{X}_1\overline{\ }$（分别地，$\overline{\ }\mathbb{X}_2$）表示 $\mathbb{B}(v_1; r)$（分别地，$\mathbb{B}(v_2; r)$）。那么，

$$
\begin{aligned}
w(x) &= \mathbb{E}\left[a(x|\overline{x})\right] \\
&= \int_{\overline{\mathbb{X}}} a(x|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\
&= \int_{\overline{\mathbb{X}}_1} a(x|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\
&= p_{\overline{\mathbb{X}}}(\overline{x}) \left\{ \int_{\overline{\mathbb{X}}_1} a(x|\overline{x})d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})d\overline{x} \right\} \qquad (p_{\mathbb{X}} \text{ 是常值函数})
\end{aligned}
$$

这里，我们考虑情况 1 和情况 2。首先，情况 1 是指 $x \in \mathbb{B}(v_1; 2r) \setminus \mathbb{B}(v_2; 2r)$ 或 $x \in \mathbb{B}(v_2; 2r) \setminus \mathbb{B}(v_1; 2r)$ 成立。由于在这种情况下，只需证明 $x \in \mathbb{B}(v_1; 2r) \setminus$

$\mathbb{B}(v_2;\,2r)$ 成立，因此我们可以假设这一条件成立。于是，$\int_{\mathbb{X}_1} a(x|\underline{x})d\underline{x} = $体积$(\mathbb{B}(v_1;r))$体积$(\mathbb{B}(v_1;\,2r))^{-1}$ 且 $\int_{\mathbb{X}_2} a(x|\underline{x})d\underline{x} = 0$。因此，$w(x) = 1/(2$体积$(\mathbb{B}(v_1;2\overline{r})))$。其次，情况 2 是指 $x \in \mathbb{B}(v_1;\,2r) \cap \mathbb{B}(v_2;\,2r)$。于是，$\int_{\mathbb{X}_1} a(x|x)dx = \int_{\mathbb{X}_2} a(\underline{x}|x)\underline{d}x =$体积$(\mathbb{B}(\underline{v_1};r))$体积$(\mathbb{B}(v_1;\,2r))^{-1}$。因此，$w(x) = 1/$体积$(\mathbb{B}(v_1;2r))$。由于 $r>0$，这意味着体积$(\mathbb{B}(v_1;2r)) > 0$。因此，$w(x) > 0$ 对两种情况都成立。

接下来，我们计算

$$
\begin{aligned}
w(x,x') &= \mathbb{E}_{\overline{x}}\left[a(x|\overline{x})a(x'|\overline{x})\right] \\
&= \int_{\overline{\mathbb{X}}} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\
&= \int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})p_{\overline{\mathbb{X}}}(\overline{x})d\overline{x} \\
&= p_{\overline{\mathbb{X}}}(\overline{x})\left\{\int_{\overline{\mathbb{X}}_1} a(x|\overline{x})a(x'|\overline{x})d\overline{x} + \int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x}\right\}. \qquad (p_{\mathbb{X}} \text{ 是一个常值函数})
\end{aligned}
$$

这里，我们考虑情况 A、情况 B、情况 C 和情况 D。首先，情况 A 是指 x 和 x' 都属于 $\mathbb{B}(v_1;2r) \setminus \mathbb{B}(v_2;\,2r)$（注意，$x$ 和 x' 都属于 $\mathbb{B}(v_2;\,2r) \setminus \mathbb{B}(v_1;\,2r)$ 时的计算相同）。接着，$\int_{\mathbb{X}_1} a(x|x)a(x'|x)dx =$体积$(\mathbb{B}(v_1;r))/$体积$(\mathbb{B}(v_1;\,2r))^2$，且 $\int_{\overline{\mathbb{X}}_2} a(x|\overline{x})a(x'|\overline{x})d\overline{x} = 0$。因此，$w(x,x') = \{2($体积$(\overline{\mathbb{B}}(v_1;\,2r)))^2\}^{-1}$。这里回想一下，$w(x) = w(x') = 1/(2$体积$(\mathbb{B}(v_1;\,2r)))$，那么我们有 $\mathsf{sim}(x,x';\lambda) = 2-\lambda$。其次，情况 B 是指 $x \in \mathbb{B}(v_1;2r) \setminus \mathbb{B}(v_2;\,2r)$ 且 $x' \in \mathbb{B}(v_2;2r) \setminus \mathbb{B}(v_1;2r)$（$x \in \mathbb{B}(v_2;\,2r) \setminus \mathbb{B}(v_1;\,2r)$ 且 $x' \in \mathbb{B}(v_1;2r) \setminus \mathbb{B}(v_2;2r)$ 时的计算相同）。接着，$\int_{\mathbb{X}_1} a(x|x)a(x'|x)dx = \int_{\mathbb{X}_2} a(x|x)a(x'|x)dx = 0$。因此，$\mathsf{sim}(x,x';\lambda) = 0-\lambda = -\lambda$。第三，情况 C 是指 x 和 x' 都属于 $\mathbb{B}(v_1;2r) \cap \mathbb{B}(v_2;2r)$。接着，$\int_{\mathbb{X}_1} a(x|x)a(x'|x)dx = \int_{\mathbb{X}_2} a(x|x)a(x'|x)dx =$体积$(\mathbb{B}(v_1;\,r))/$体积$(\mathbb{B}(v_1;\,2r))^2$。由于 $w(x) = w(x') = 1/$体积$(\mathbb{B}(v_1;\,2r))$，所以 $\mathsf{sim}(x,x';\lambda) = 1-\lambda$。最后，在情况 D 中，考虑其他情况并集的补集。根据设定，我们可以假设 x 属于 $\mathbb{B}(v_1;2r) \cap \mathbb{B}(v_2;\,2r)$，而 x' 属于 $\mathbb{B}(v_1;2r) \setminus \mathbb{B}(v_2;2r)$。接着，$\int_{\mathbb{X}_1} a(x|x)a(x'|x)dx = $体积$(\mathbb{B}(v_1;r))/$体积$(\mathbb{B}(v_1;\,2r))^2$，且 $\int_{\mathbb{X}_2} a(x|x)a(x'|x)dx = 0$。由于 $w(x) = 1/$体积$(\mathbb{B}(v_1;\,2r))$ 且 $w(x') = 1/(2$体积$(\mathbb{B}(v_1;\,2r)))$，我们有 $\mathsf{sim}(x,x';\lambda) = 1-\lambda$。因此，

$$
\mathsf{sim}(x,x';\lambda) = \begin{cases} 2-\lambda, & \text{if Case A holds,} \\ -\lambda, & \text{if Case B holds,} \\ 1-\lambda, & \text{if Case C holds,} \\ 1-\lambda, & \text{if Case D holds.} \end{cases}
$$

最后，取 $\delta \in (-\lambda, 1-\lambda]$。然后，根据上文对 $\mathsf{sim}(x,x';\lambda)$ 的计算，假设 2 中的条件均得到满足。

□

F.2 重新审视 SSL-HSIC

Li 等人（2021年）提出了名为 SSL-HSIC 的框架，该框架基于希尔伯特-施密特独立性准则（HSIC，（Smola 等人，2007年））定义。他们证明，在某些条件下，对于表示特征向量（或标签）的随机变量 Z（分别为 Y），可以得到

$$
\mathrm{HSIC}(Z,Y) = c\left(\mathbb{E}_{x,x^+}\left[k(f(x),f(x^+))\right] - \mathbb{E}_{x,x^-}\left[k(f(x),f(x^-))\right]\right),
$$

其中 $c > 0$。Li 等人（2021年）将 SSL-HSIC 的损失定义为：

$$L_{\text{SSL-HSIC}}(f;\kappa) = -\text{HSIC}(Z,Y) + \kappa\sqrt{\text{HSIC}(Z,Z)},$$

其中 $\kappa \in \mathbb{R}$。

在 $\kappa > 0$ 的情况下，我们有

$$L_{\text{KCL}}(f;1) \lesssim L_{\text{SSL-HSIC}}(f;\kappa).$$

第 F.3 节补充信息3

F.3.1 InfoNCE 的变体之间的关系

我们首先定义 InfoNCE 的变体（Chen 等人，2020a；van den Oord 等人，2018年）：

- 解耦 InfoNCE 损失，这是 Chen 等人提出的解耦 NT-Xent 损失的一个变体（Chen 等人（2021年）：

$$\widetilde{L}_{\text{NCE}}(f;\tau,\lambda) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \lambda\mathbb{E}_{\substack{x,x^+\\{x_i^-\}}}\left[\log\left(e^{\frac{f(x)^\top f(x^+)}{\tau}} + \sum_{i=1}^{M} e^{\frac{f(x)^\top f(x_i^-)}{\tau}}\right)\right].$$

- 对比损失的渐近形式（Wang 和 Isola，2020年）按照 Chen 等人（2021年）的方式进行解耦：

$$\widetilde{L}_{\infty\text{-NCE}}(f;\tau,\lambda) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \lambda\mathbb{E}_{x}\left[\log\mathbb{E}_{x'}\left[e^{\frac{f(x)^\top f(x')}{\tau}}\right]\right],$$

- InfoNCE 损失是解耦对比学习损失的一种变体（Yeh 等人，2022年）：

$$\widetilde{L}_{\text{NCE}}(f;\tau,1) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \mathbb{E}_{x,\{x_i^-\}}\left[\log\left(\sum_{i=1}^{M} e^{\frac{f(x)^\top f(x_i^-)}{\tau}}\right)\right].$$

- 按照 Chen 等人（2021年）的做法，InfoNCE 损失是带有额外权重参数的解耦对比学习损失的一种变体：

$$\widetilde{L}_{\text{NCE}}(f;\tau,\lambda) = -\mathbb{E}_{x,x^+}\left[\frac{f(x)^\top f(x^+)}{\tau}\right] + \lambda\mathbb{E}_{x,\{x_i^-\}}\left[\log\left(\sum_{i=1}^{M} e^{\frac{f(x)^\top f(x_i^-)}{\tau}}\right)\right].$$

请注意，$\widetilde{L}_{\text{NCE}}(f;\tau)$ 和 $\widetilde{L}_{\infty\text{-NCE}}(f;\tau)$ 在第2节中分别与 $L_{\text{NCE}}(f;\tau,1)$ 以及 $L_{\infty\text{-NCE}}(f;\tau,1)$ 一致。我们证明以下事实：

命题 5. 以下关系成立：

$$\tau^{-1}L_{\text{LinKCL}}(f;\lambda) \le \widetilde{L}_{\text{NCE}}(f;\tau,\lambda) + \lambda\log M^{-1}, \tag{29}$$

$$\tau^{-1}L_{\text{LinKCL}}(f;\lambda) \le \widetilde{L}_{\infty\text{-NCE}}(f;\tau,\lambda), \tag{30}$$

$$\tau^{-1}L_{\text{LinKCL}}(f;\lambda) \le \widetilde{L}_{\text{NCE}}(f;\tau,\lambda) + \lambda\log M^{-1}. \tag{31}$$

证明. 根据 $L_{\mathrm{NCE}}(f;\tau,\lambda)$ 的定义，我们有

$$
\begin{aligned}
&\widetilde{L}_{\mathrm{NCE}}(f;\tau,\lambda)+\lambda\log\frac{1}{M}\\
&=-\tau^{-1}\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right]+\lambda\mathbb{E}_{x,x^+,\{x_i^-\}}\left[\log\left(\frac{1}{M}e^{f(x)^\top f(x^+)/\tau}+\frac{1}{M}\sum_{i=1}^{M}e^{f(x)^\top f(x_i^-)/\tau}\right)\right]\\
&\geq-\tau^{-1}\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right]+\lambda\mathbb{E}_{x,\{x_i^-\}}\left[\log\left(\frac{1}{M}\sum_{i=1}^{M}e^{f(x)^\top f(x_i^-)/\tau}\right)\right]\\
&\geq-\tau^{-1}\mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right]+\tau^{-1}\lambda\mathbb{E}_{x,\{x_i^-\}}\left[\frac{1}{M}\sum_{i=1}^{M}f(x)^\top f(x_i^-)\right]\\
&=\tau^{-1}L_{\mathrm{LinKCL}}(f;\lambda),
\end{aligned}
$$

其中，在第一个不等式中，我们利用了以下事实： $M^{-1}e^{f(x)^\top f_{(x^+)}/\tau}\geq 0$ 对于任意 x， $x^+\in\mathbb{X}$而在第二个不等式中，我们使用了詹森不等式。注意，当 $\lambda=1$时，我们得到（1）。

（31）的证明与（29）的证明几乎相同。等式（30）是通过应用詹森不等式得到的。

□

F.3.2 与 SCL 的关系

我们将二次核对比损失定义为:

$$
L_{\mathrm{QKCL}}(f;\lambda)=-\mathbb{E}_{x,x^+}\left[\left(f(x)^\top f(x^+)\right)^2\right]+\lambda\mathbb{E}_{x,x^-}\left[\left(f(x)^\top f(x^-)\right)^2\right].
$$

谱对比损失 $L_{\mathrm{SCL}}(f)$（HaoChen 等人，2021年）定义如下:

$$
L_{\mathrm{SCL}}(f)=-2\mathbb{E}_{x,x^+}[f(x)^\top f(x^+)]+\mathbb{E}_{x,x^-}[(f(x)^\top f(x^-))^2]. \tag{32}
$$

以下命题是一个初等结果。

命题 6。我们有，

$$
L_{\mathrm{QKCL}}(f;2^{-1})\leq\frac{1}{2}L_{\mathrm{SCL}}(f)+\frac{1}{4}.
$$

证明。由于对每个 $t\in\mathbb{R}$ 都有 $t^2+1/4\geq t$，我们得到所要证明的结论。 □

F.4 我们的假设2 与HaoChen 和 Ma(2023年)2023年论文中的假设 3 的比较

令 $\mathbb{M}$ 为 $\mathbb{X}$的可测子集，并令 $g:\mathbb{X}\to\mathbb{R}$ 为一个函数。 HaoChen 和 Ma (2023年) 引入了以下概念，用于量化簇的内部连通性（见 HaoChen 和 Ma (2023年) 中的 (4)）：

$$
Q_{\mathbb{M}}(g):=\frac{\mathbb{E}_{x,x^+}[(g(x)-g(x^+))^2|\mathbb{M}\times\mathbb{M}]}{\mathbb{E}_{x,x^-}[(g(x)-g(x^-))^2|\mathbb{M}\times\mathbb{M}]}.
$$

这里，上述期望定义为

$$\mathbb{E}_{x,x^+}[(g(x)-g(x^+))^2|\mathbb{M}\times\mathbb{M}] = \int_{\mathbb{X}\times\mathbb{X}}(g(x)-g(x'))^2 P_+(dx,dx'|\mathbb{M}\times\mathbb{M}),$$
$$\mathbb{E}_{x,x^-}[(g(x)-g(x^-))^2|\mathbb{M}\times\mathbb{M}] = \int_{\mathbb{X}}\int_{\mathbb{X}}(g(x)-g(x'))^2 P_{\mathbb{X}}(dx|\mathbb{M})P_{\mathbb{X}}(dx'|\mathbb{M}),$$

其中我们使用记号 $dP_+ = w(x,x')d\nu_{\mathbb{X}}^{\otimes 2}$ 。我们关注以下概念，其中 HaoChen 和 Ma (2023年) 将他们的子集记为 $\{S_1,\cdots,S_m\}$。

假设 4 ($\mathcal{F}$-可实现的簇内连接大于 β，引自HaoChen 和 Ma (2023年))。 对于任意函数 $f\in\mathcal{F}$ 和任意线性头 $w\in\mathbb{R}^k$，令函数 $g(x) = w^\top f(x)$。对于任意 $i\in[m]$，我们有：

$$Q_{S_i}(g)\geq\beta.$$

总之，HaoChen 和 Ma (2023年)的假设 3与我们工作的假设 2之间的关系如下：

命题 7。假设假设 2成立。取 $\delta\in\mathbb{R}$、 $K\in\mathbb{N}$以及 $\mathbb{M}_1$、$\cdots$、$\mathbb{M}_K$，使得条件 *(A)* 和 *(B)* 得到满足。另假设：*1*）存在某个 $c>0$，使得对于每个 $i\in[K]$， $c\cdot P_+(\mathbb{M}_i\times\mathbb{M}_i)\leq P_{\mathbb{X}}(\mathbb{M}_i)^2$ 成立；*2*） $\delta+\lambda\geq 0$ 成立。则函数类 $\widetilde{\mathcal{F}}$（包含所有从 $\mathbb{X}$ 到 $\mathbb{S}^{d-1}$ 的映射）满足 *HaoChen* 和 *Ma*（*2023*年）中的假设 *3*。

证明。 任取 $f\in\widetilde{\mathcal{F}}$ 和 $w\in\mathbb{R}^d$。对于每个 $i\in[K]$，对任意 x， $x'\in\mathbb{M}_i$ 我们有$w(x,\ x')\geq(\delta+\lambda)w(x)w(x')$。由于 $\delta+\lambda\geq 0$,

$$\int_{\mathbb{M}_i\times\mathbb{M}_i}(g(x)-g(x'))^2 w(x,x')\nu_{\mathbb{X}}^{\otimes 2}(dx,dx')\geq(\delta+\lambda)\int_{\mathbb{M}_i\times\mathbb{M}_i}(g(x)-g(x'))^2 w(x)w(x')\nu_{\mathbb{X}}^{\otimes 2}(dx,dx').$$

这里，利用 $c\cdot P_+(\mathbb{M}_i\times\mathbb{M}_i)\leq P_{\mathbb{X}}(\mathbb{M}_i)^2,$，我们有

$$\begin{aligned}&\frac{1}{P_+(\mathbb{M}_i\times\mathbb{M}_i)}\int_{\mathbb{M}_i\times\mathbb{M}_i}(g(x)-g(x'))^2 w(x,x')\nu_{\mathbb{X}}^{\otimes 2}(dx,dx')\\ \geq\ & c(\delta+\lambda)\frac{1}{P_{\mathbb{X}}(\mathbb{M}_i)^2}\int_{\mathbb{M}_i\times\mathbb{M}_i}(g(x)-g(x'))^2 w(x)w(x')\nu_{\mathbb{X}}^{\otimes 2}(dx,dx').\end{aligned}$$

上述不等式意味着，

$$\int_{\mathbb{X}\times\mathbb{X}}(g(x)-g(x'))^2 P_+(dx,dx'|\mathbb{M}_i\times\mathbb{M}_i)\geq c(\delta+\lambda)\int_{\mathbb{X}}\int_{\mathbb{X}}(g(x)-g(x'))^2 P_{\mathbb{X}}(dx|\mathbb{M}_i)P_{\mathbb{X}}(dx'|\mathbb{M}_i).$$

因此，我们得到 $Q_{\mathbb{M}_i}(g)\geq c(\delta+\lambda)$，其中 $g(x) = w^\top f(x)$。 □

上述命题表明，在这些假设下，对于任意 $i\in[K]$ 和满足 $f\in\widetilde{\mathcal{F}}$的 $g(x) = w^\top f(x)$，其内部连通性 $Q_{\mathbb{M}_i}(g)$ 的下界为 $c(\delta+\lambda)$。因此，如果 $c\cdot P_+(\mathbb{M}_i\times\mathbb{M}_i)\leq P_{\mathbb{X}}(\mathbb{M}_i)^2$ （对每个 $i\in[K]$均成立）且 $\delta+\lambda\geq 0$ 成立，那么我们工作的假设 2 是 HaoChen 和 Ma（2023）中假设 3 的充分条件。

注释 4. 我们可以显式构造上述陈述中的一个正值 c 。在本注释中，我们展示一种简单的构造方法。令 X, Y为概率空间 $(\Omega\ P_\Omega)$ 上的随机变量，其联合概率分布为 P_+ ，边缘分布为 $P_{\mathbb{X}}$。记 $p_1 = P_\Omega(X \in \mathbb{M}_i)$ 和$p = P_\Omega(X \in \mathbb{M}_i, Y \in \mathbb{M}_i)$。此处，令 V为随机变量 $\mathbb{1}_{\{X\in\mathbb{M}_i\}}$和 $\mathbb{1}_{\{Y\in\mathbb{M}_i\}}$的协方差矩阵。 V的半正定性意味着不等式 $(p_1 - p_1^2)^2 - (p - p_1^2)^2 \geq 0$。当 $2p_1^2 - p_1 \leq p \leq p_1$时，该不等式成立。结合这一事实与性质 $p \geq 0$，我们得到

$$\max\{2p_1^2 - p_1, 0\} \leq p \leq p_1,$$

这意味着，

$$P_{\mathbb{X}}(\mathbb{M}_i) \cdot \max\{2P_{\mathbb{X}}(\mathbb{M}_i) - 1, 0\} \cdot P_+(\mathbb{M}_i \times \mathbb{M}_i) \leq P_{\mathbb{X}}(\mathbb{M}_i)^2.$$

因此，如果 $P_{\mathbb{X}}（\mathbb{M}_i）\ > 1/2$ 对每个 $i \in [K]$都成立，那么我们可以取

$$c = \min_{i\in[K]}\{P_{\mathbb{X}}(\mathbb{M}_i)(2P_{\mathbb{X}}(\mathbb{M}_i) - 1)\}.$$

F.关于第 5.3 节泛化界的进一步讨论5.3

本节讨论我们的泛化误差界与其他关于对比学习的研究结果之间的差异。我们将这些差异总结如下。

- Arora 等人（2019）、Ash 等人（2022）、Lei 等人（2023）以及 Zou 和 Liu（2023）考虑了这样一种情况：对于一对 (x, x^+)，存在与其他随机变量独立的 M元组样本 $(x_1^-, \cdots, x_M^-)$。因此，我们的问题设定与这些研究不同。尤其是，在我们的分析中，还需要处理 X_i, X_i'，$i \in [n]$ 不一定相互独立的情况，而标准技术（例如，参见 Mohri 等人（2018））无法适用。我们转而采用了 Zhang 等人（2019）给出的关于相关随机变量的 McDiarmid 不等式结果。

- Zhang 等人 所考虑的经验损失(2022年) 的定义方式与我们的经验核对比损失不同。此外，我们的证明技术也不同于 Zhang 等人(2022年)。

- HaoChen 等人 (2021年)；Nozawa 等人 (2020年) 研究了增强样本不一定相互独立的情形。Nozawa 等人 (2020 年) 运用 PAC- 贝叶斯界理论（ Guedj，2019 年），而 HaoChen 等人 (2021年) 给出了高概率界。我们的分析与 Nozawa 等人(2020年) 不同，因为我们的分析基于若干集中不等式。 HaoChen 等人 (2021年) 研究了由原始样本定义的经验谱对比损失，该损失以根据给定原始样本的条件分布抽取的增强样本为依据来表示期望（参见 HaoChen 等人 (2021年) 中的第 4.1 节）。另一方面，我们为仅由增强样本定义的经验核对比损失推导了泛化误差界。

- Wang 等人(2022b) 建立了谱对比损失（HaoChen 等人，2021年）的泛化误差界，其中他们的分析提高了 HaoChen

等人 (2021年）。在他们的分析中，他们以不同于我们的方法分解谱对比损失的第二项（有关其分解的详细信息，请参见 Wang 等人 (2022b）的命题 D.1 的证明）。此外，他们利用了 Cl´emen¸con 等人 (2008年）所给出的集中不等式（参见 Wang 等人 (2022b）中的公式（51））， 而我们使用的是 Zhang 等人 (2019年）证明的结果。因此，我们在定理6 的证明中使用的技术不同于 Wang 等人(2022b）所使用的技术。

F.6 与 Robinson 等（2021b）的详细比较

Robinson 等人（2021b）从理论和实证两个角度探讨了对比学习中的困难负样本采样问题。他们还通过引入 1-NN 分类器，为困难负样本目标建立了泛化界（关于 1-NN 分类器的定义，参见 Robinson 等人（2021b）中的定理 5 陈述）。我们将本研究的结果与 Robinson 等人（2021b）的理论分析进行了详细比较。主要区别如下：

- Robinson 等人（2021b）的理论结果所采用的问题设定基于 Arora 等人（2019年）的问题设定，即依赖于条件独立性假设。另一方面，我们并不依赖该假设，而是采用相似度函数 $\mathrm{sim}(\cdot,\ \cdot;\ \lambda)$。

- 在 Robinson 等（2021b）的定理 5 证明中（另见定理 8 及其论文中的证明），监督损失以上界项 $\mathbb{E}_c\mathbb{E}_{x,x^+\sim_{iid}p(\cdot|c)}\|f(x) - f(x^+)\|^2$ 为界。总之，Robinson 等（2021b）的定理 5 与我们工作的定理 2 之间的区别如下：（i）在 Robinson 等（2021b）的定理 8 证明中的上界分子里，出现了上述项。另一方面，在我们工作的定理 2 中，出现的是量 $\alpha(f)$。（ii）我们的上界包含量 $\Delta\min(f)$。（iii）我们还注意到，我们在定理 2 中采用的证明技术不同于 Robinson 等（2021b）所采用的技术。

- 请注意，Robinson 等人(2021b) 的分析采用的标签是一个随机变量，而我们的分析采用的是确定性标注函数。

F.7 与 Huang 等人(2023年)；Zhao 等人(2023年) 的详细比较)

Huang 等人(2023年) 给出了利用 1-NN 分类器的泛化界（关于 Huang 等人(2023年) 提出的 1-NN 分类器定义，参见其论文第 2 节）。此外，Zhao 等人(2023年) 扩展了 Huang 等人(2023年) 的结果。因此，有必要讨论 Huang 等人(2023年)；Zhao 等人(2023年) 的结果与我们的定理 2) 之间的差异。下面总结这些差异：

- Huang 等人（2023年）证明，如果特征空间中的簇中心彼此足够远（注意，他们将其称为散度），那么其监督误差函数可由对齐 项乘以若干常数和参数来上界。因此，他们的结果并未表明散度 与监督误差直接相关，即散度 项并未出现在其监督误差的上界中。另一方面

，我们证明，与再生核希尔伯特空间中的散度 相关的量也有助于给出监督误差的上界（见定理2）。

• 在Huang 等人（2023年）；Zhao 等人（2023年）中，很少有人研究他们的结果适用于多大范围的编码器模型。另一方面，我们的定理2 仅要求属于 $\mathcal{F}$的编码器具有实质意义（定义2）。特别地，假设 k 是线性核，那么从这个意义上说，我们的研究中的定理2 细化了Huang 等人（2023年）提出的定理1。

• Huang 等人（2023年）；Zhao 等人（2023年）采用了称为（σ， δ）-增广的概念，而我们的分析则基于相似度函数 sim(·， ·； λ) 的定义，采用假设2 。

• Huang 等人（2023年）；Zhao 等人（2023年）通常假设编码器 f 是利普希茨函数。与此同时，我们的主要结果并不要求 f 必须是利普希茨函数。

•Zhao 等人（2023年）考虑了下游分类任务的平方损失（见其论文中的定理 3.2）。另一方面，我们考虑的是分类误差。

G KCL 与归一化割之间的联系 t

在本节中，我们给出第4.1节的补充信息。贯穿本节，我们假设 $\inf_{x\in\mathbb{X}} w(x) < \infty$ 和 $\sup_{x\in\mathbb{X}}\sup_{x\in\mathbb{X}} a(x|x) < \infty$ 成立。注意，假设 $\sup_{x\in\mathbb{X}}\overline{\sup}_{x\in\mathbb{X}} a(x|x) < \infty$ 蕴含 $\sup_{x\in\mathbb{X}} w(x) < \infty$。

G.1 归一化割的问题设定

在本节中，我们首先基于 Shi 和 Malik(2000)；Terada 和 Yamamoto (2019年)；Von Luxburg (2007年)解释总体归一化割问题。假设 K 中总共有 $\mathbb{X}$个簇。遵循 Terada 和 Yamamoto (2019年)，总体层面的归一化割的优化问题表述如下：

$$\min_{\mathbb{V}_1,\cdots,\mathbb{V}_K} \sum_{i=1}^{K} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)} \tag{33}$$

其中，上述问题中的最小值是在所有可能的 K 个两两不相交的非空可测子集 $\mathbb{V}_1$、$\cdots$、$\mathbb{V}_K$的组合上取得的，它们满足$\bigcup_{i=1}^{K}\mathbb{V}_i = \mathbb{X}$，并且 W和 vol(·) 定义如下：

$$W(\mathbb{V}_i, \mathbb{V}_i^c) = \int_{(x,x')\in\mathbb{V}_i\times\mathbb{V}_i^c} \mathrm{sim}(x, x'; \lambda) w(x) w(x') d\nu_{\mathbb{X}}^{\otimes 2}(x, x'), \tag{34}$$

$$\mathrm{vol}(\mathbb{V}_i) = \int_{\mathbb{V}_i} w(x) d\nu_{\mathbb{X}}(x), \tag{35}$$

其中 $\nu_{\mathbb{X}}^{\otimes 2} := \nu_{\mathbb{X}} \otimes \nu_{\mathbb{X}}$ 是乘积测度。这里还要注意，vol(·) 是集合$\mathbb{V}_i$的体积。 Terada 和 Yamamoto(2019年) 考虑了将再生核用作相似度的情况

度量：参见 Terada 和 Yamamoto (2019年) 中的定理 7。注意，一些已有研究处理了可测划分问题，例如比率割和 Cheeger 割（Trillos 等人，2016）。

记 $L^2(\mathbb{X},\ P_{\mathbb{X}})$ 为域 $\mathbb{R}$ 上的希尔伯特空间，其中包含定义在 $\mathbb{X}$ 上、关于 $P_{\mathbb{X}}$-几乎处处实值且平方可积的函数，其内积为 $\langle f,g\rangle_{L^2(\mathbb{X},\ P_{\mathbb{X}})} =\int f(x)g(x)dP_{\mathbb{X}}(x)$。令 $U\colon \mathbb{R}^K \to L^2(\mathbb{X},\ P_{\mathbb{X}})$ 为如下定义的线性算子，

$$(Uz)(\cdot)=\sum_{i=1}^{K}\frac{\mathbb{1}_{\mathbb{V}_i}(\cdot)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}}z_i,\quad z=(z_1,\cdots,z_K)^\top, \tag{36}$$

其中，若 $x\in\mathbb{V}_i$，则 $\mathbb{1}_{\mathbb{V}_i}(x)=1$；若 $x\notin\mathbb{V}_i$，则其值为 0；并且，对每个 $i\in[K]$，$z_i:=\langle z,e_i\rangle_{\mathbb{R}^K}$ 都具有 $\mathbb{R}^K$ 的一组正交归一基 $\{e_i\}_{i=1}^K$。注意，在 $|\mathbb{X}|<\infty$ 的设定下，线性算子 U 等于 $(\mathbb{1}_{\mathbb{V}_j}(x_i)/\sqrt{\text{体积}(\mathbb{V}_j)})_{ij}$。因此，$U$ 的定义与归一化割的经典理论（Shi 和 Malik，2000）一致。此外，每个增强数据 $x\in\mathbb{X}$ 都属于子集 $\mathbb{V}_1$、$\cdots$、$\mathbb{V}_K$ 之一。由于 $\mathbb{V}_1$、$\cdots$、$\mathbb{V}_k$ 两两不相交，线性算子 U 有界，且伴随算子 $U^\dagger$ 唯一存在。这里，Von Luxburg（2007）解释说，归一化割问题的目标函数可以改写为一个组合优化问题。将 Von Luxburg（2007）提出的论证应用于我们的设定，可得

$$\sum_{i=1}^{K}\frac{W(\mathbb{V}_i,\mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)}=-\mathrm{Tr}(U^\dagger AU)+(1-\lambda)K, \tag{37}$$

其中 $A\colon L^2(\mathbb{X},\ P_{\mathbb{X}})\to L^2(\mathbb{X},\ P_{\mathbb{X}})$ 是一个定义如下的希尔伯特-施密特积分算子

$$A\psi(\cdot)=\int \mathrm{sim}(\cdot,x;\lambda)\psi(x)w(x)d\nu_{\mathbb{X}}(x)\quad \psi\in L^2(\mathbb{X},P_{\mathbb{X}}), \tag{38}$$

且 $-\mathrm{Tr}(U^\dagger AU)=-\sum_{i=1}^K\langle U^\dagger AUe_i,e_i\rangle_{\mathbb{R}^K}$。式 (37) 的证明与 Von Luxburg(2007年) 的证明十分相似；在附录 G.3 中，我们给出了一个扩展版本的证明。以下命题表明 A 是定义良好的。

命题 8。 假设第 *2.1* 节所述的设定成立，$\inf_{x\in\mathbb{X}}w(x)>0$，且$\sup_{x\in\mathbb{X}}\sup_{x\in\mathbb{X}}a(x|x)<\infty$ 成立。那么，$^-$积分算子 $\bar{A}$定义良好。

证明。 我们可以计算

$$\begin{aligned}\left|\int \mathrm{sim}^2(x,x;\lambda)w(x)d\nu_{\mathbb{X}}(x)\right| &= \left|\int\left(\frac{w(x,x)}{w(x)w(x)}-\lambda\right)^2 w(x)d\nu_{\mathbb{X}}(x)\right| \\ &\le \int\left(\left(\frac{w(x,x)}{w(x)w(x)}\right)^2+2\lambda\frac{w(x,x)}{w(x)w(x)}+\lambda^2\right)w(x)d\nu_{\mathbb{X}}(x) \\ &< +\infty,\end{aligned}$$

其中我们使用了以下假设：$\inf_{x\in\mathbb{X}}w(x)>0$，$\sup_{x\in\mathbb{X}}\sup_{x\in\mathbb{X}}a(x|x)<\infty$。 □

假设与核函数 $\mathcal{H}_k$ 相关联的再生核希尔伯特空间 k 的维数大于或等于 K。在下一节中，为方便起见，将（33）重新定义为：

$$\min_{\mathbb{V}_1,\cdots,\mathbb{V}_K}\sum_{i=1}^{\infty}\frac{W(\mathbb{V}_i,\mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)},$$

其中，我们对每个 $\mathbb{V}_j = \emptyset$ 定义 $j > K$。此外，将（36）重新定义为线性算子 $U:\ \mathcal{H}_k \to L^2$（$\mathbb{X}$，$P_{\mathbb{X}}$）。

$$(U\psi)(\cdot) = \sum_{i=1}^{\infty} \frac{\mathbb{1}_{\mathbb{V}_i}(\cdot)}{\sqrt{\text{vol}(\mathbb{V}_i)}} \langle \psi, e_i \rangle_{\mathcal{H}_k},$$

其中 $\{e_j\}_{j=1}^{\infty}$ 是 $\mathcal{H}_k$ 的正交归一基（如果 $\mathcal{H}_k$ 是有限维的，则我们认为$\{e_j\}_{j=1}^{\infty}$ 仅包含有限个非零元素），并定义 $\mathbb{1}_{\mathbb{V}_i}(\cdot)/\sqrt{\text{体积}(\mathbb{V}_i)} = 0$ 对每个$i > K$ 作为记号。于是，我们得到以下与（37）类似的恒等式：

$$\sum_{i=1}^{\infty} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\text{vol}(\mathbb{V}_i)} = -\text{Tr}(U^\dagger A U) + (1-\lambda)K. \tag{39}$$

为完整起见，我们给出附录 39 中恒等式（G.3）的证明。

G.2 通过再生核希尔伯特空间连接 KCL 与归一化割

令 k：$\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ 为一个连续、对称且正定的核函数，其再生核希尔伯特空间$\mathcal{H}_k$ 是一个 K 维希尔伯特空间（K 为有限或 ∞）；关于再生核理论，参见例如 Aronszajn（1950年）；Berlinet 和 Thomas-Agnan（2004年）；Steinwart and Christmann（2008年）。许多核函数满足这些条件，例如高斯核、多项式核和线性核。由于 $\mathbb{S}^{d-1}$ 是可分的，再生核希尔伯特空间 $\mathcal{H}_k$ 具有至多可数的正交归一基（例如，参见 Berlinet 和 Thomas-Agnan（2004年））。令 $\{e_j\}_{j=1}^{\infty}$ 为 $\mathcal{H}_k$的一个可数正交归一基，其中，如果 $\mathcal{H}_k$ 是有限维的，$\{e_j\}_{j=1}^{\infty}$ 仅包含有限个非零元素。注意，无论选择 $\{e_j\}_{j=1}^{\infty}$为何，我们的构造均有效。回顾第 2.1 节中的问题设定。于是，线性算子 $H:\ \mathcal{H}_k \to L^2(\mathbb{X}, P_{\mathbb{X}})$ 定义为

$$(H\varphi)(\cdot) = \langle h(f(\cdot)), \varphi \rangle_{\mathcal{H}_k}, \tag{40}$$

对于 $\varphi \in \mathcal{H}_k$。令 $\|\cdot\|_{\mathcal{H}_k}$ 为再生核希尔伯特空间 $\mathcal{H}_k$的范数。那么，对于在（40）中定义的线性算子 H，有如下结论：

命题 9。线性算子 H 定义良好，即对每个 $\varphi \in \mathcal{H}_k$，$H\varphi \in L^2$（$\mathbb{X}$，$P_{\mathbb{X}}$）。此外，$H$ 是连续的，即对于强收敛到 φ （属于 $\mathcal{H}_k$）的序列φ_j，有 $H\varphi_j$ 收敛到 $H\varphi$。

证明。对于任意 $\varphi \in \mathcal{H}_k$，我们有

$$\begin{aligned}
\int |(H\varphi)(x)|^2 w(x) d\nu_{\mathbb{X}}(dx) &= \int |\langle h(f(x)), \varphi \rangle_{\mathcal{H}_k}|^2 w(x) d\nu_{\mathbb{X}}(x) \\
&\le \int \|h(f(x))\|_{\mathcal{H}_k}^2 \|\varphi\|_{\mathcal{H}_k}^2 w(x) d\nu_{\mathbb{X}}(x) \\
&\le \|\varphi\|_{\mathcal{H}_k}^2 \mathbb{E}_x \left[k(f(x), f(x))\right] < +\infty.
\end{aligned}$$

这里在第二个不等式中，我们使用柯西-施瓦茨不等式；在最后一个等式中，我们利用了 $\mathbb{S}^{d-1}$ 是紧的且 k 是连续的这一事实。此外，如果 $\varphi_j \to \varphi$ 以强收敛的意义

在 $\mathcal{H}_k$ 中，那么我们有

$$\begin{aligned}\|\langle h(f(\cdot)),\varphi_j\rangle_{\mathcal{H}_k}-\langle h(f(\cdot)),\varphi\rangle_{\mathcal{H}_k}\|^2_{L^2(\mathbb{X},P_{\mathbb{X}})} &= \int |\langle h(f(x)),\varphi_j-\varphi\rangle_{\mathcal{H}_k}|^2 w(x)d\nu_{\mathbb{X}}(x)\\ &\le \|\varphi_j-\varphi\|^2_{\mathcal{H}_k}\mathbb{E}_x\left[k(f(x),f(x))\right]\\ &\longrightarrow 0 \quad (j\to\infty).\end{aligned}$$

因此， $H\varphi_j$ 收敛到 $H\varphi$，证明至此结束。 □

命题 9 表明 H 有界。因此，伴随算子 $H^\dagger: L^2(\mathbb{X}, P_{\mathbb{X}}) \to \mathcal{H}_k$唯一存在。

现在回顾相似度函数 sim(·, ·; λ) 在固定 λ 下（2），并考虑使用线性算子 H （40）对组合问题（33）进行松弛，其中定义如下：我们将（ U 37）中的线性算子替换为 H，由此得到目标函数–$\mathrm{Tr}(H^\dagger AH)$。于是，以下命题成立。

命题 10. 我们有

$$-\mathrm{Tr}(H^\dagger AH) = -\mathbb{E}_{x,x^+}\left[k(f(x),f(x^+))\right] + \lambda\mathbb{E}_{x,x^-}\left[k(f(x),f(x^-))\right].$$

证明。 根据 sim(x, x'; λ) 的定义，

$$\begin{aligned}(A\psi)(x) &= \int \mathrm{sim}(x,x';\lambda)\psi(x')w(x')d\nu_{\mathbb{X}}(x')\\ &= \int\left(\frac{w(x,x')}{w(x)w(x')}-\lambda\right)\psi(x')w(x')d\nu_{\mathbb{X}}(x')\\ &= \underbrace{\int\frac{w(x,x')}{w(x)w(x')}\psi(x')w(x')d\nu_{\mathbb{X}}(x')}_{:=(A_{\mathrm{pos}}\psi)(x)} - \lambda\underbrace{\int\psi(x')w(x')d\nu_{\mathbb{X}}(x')}_{:=(A_{\mathrm{neg}}\psi)(x)}.\end{aligned}$$

首先，让我们证明恒等式

$$\mathrm{Tr}(H^\dagger A_{\mathrm{pos}}H) = \mathbb{E}_{x,x^+}\left[k(f(x),f(x^+))\right].$$

证明如下：根据 H 的定义，

$$(He_i)(x) = \langle h(f(x)),e_i\rangle_{\mathcal{H}_k} \quad x\in\mathbb{X}.$$

于是有

$$\begin{aligned}(A_{\mathrm{pos}}He_i)(x) &= \int\frac{w(x,x')}{w(x)w(x')}w(x')\langle h(f(x')),e_i\rangle_{\mathcal{H}_k}d\nu_{\mathbb{X}}(x')\\ &= \int\frac{w(x,x')}{w(x)}\langle h(f(x')),e_i\rangle_{\mathcal{H}_k}d\nu_{\mathbb{X}}(x').\end{aligned}$$

H这里，伴随算子 $H^\dagger$ 满足以下恒等式；对于 $\psi\in L^2$(𝕏, $P_{\mathbb{X}}$),

$$\langle He_i,\psi\rangle_{L^2(\mathbb{X},P_{\mathbb{X}})} = \langle e_i,H^\dagger\psi\rangle_{\mathcal{H}_k}.$$

利用这一关系可得以下表示:

$$
\begin{aligned}
&H^\dagger A_{\mathrm{pos}} H e_j \\
=&\sum_{i=1}^{\infty} \langle H^\dagger A_{\mathrm{pos}} H e_j, e_i \rangle_{\mathcal{H}_k} e_i \\
=&\sum_{i=1}^{\infty} \langle A_{\mathrm{pos}} H e_j, H e_i \rangle_{L^2(\mathbb{X}, P_{\mathbb{X}})} e_i \\
=&\sum_{i=1}^{\infty} \left(\int w(x) \langle h(f(x)), e_i \rangle_{\mathcal{H}_k} \int \frac{w(x,x')}{w(x)} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x) \right) e_i \\
=&\sum_{i=1}^{\infty} \left(\int \int w(x,x') \langle h(f(x)), e_i \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \right) e_i.
\end{aligned}
$$

因此,

$$
\begin{aligned}
\mathrm{Tr}(H^\dagger A_{\mathrm{pos}} H) &= \sum_{j=1}^{\infty} \langle H^\dagger A_{\mathrm{pos}} H e_j, e_j \rangle_{\mathcal{H}_k} \\
&= \sum_{j=1}^{\infty} \int \int w(x,x') \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&= \int \int w(x,x') \sum_{j=1}^{\infty} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&= \int \int w(x,x') \langle h(f(x)), h(f(x')) \rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\
&= \mathbb{E}_{x,x^+} \left[k(f(x), f(x^+)) \right].
\end{aligned}
$$

注意到，上述第三个等式源于控制收敛定理。事实上，级数 $\sum_{j=1}^{n} \langle h(f(x))\ e_j \rangle_{\mathcal{H}_k} \langle h(f(x'))\ e_j \rangle_{\mathcal{H}_k}$ 在 $\mathbb{X} \times \mathbb{X}$ 上逐点收敛于 $\langle h(f(x))\ h(f(x)) \rangle_{\mathcal{H}_k}$，并且

$$
\begin{aligned}
\left| \sum_{j=1}^{n} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} \right| &\leq \sum_{j=1}^{n} \left| \langle h(f(x)), e_j \rangle_{\mathcal{H}_k} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k} \right| \\
&\leq \left(\sum_{j=1}^{n} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \left(\sum_{j=1}^{n} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{\infty} \langle h(f(x)), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \left(\sum_{j=1}^{\infty} \langle h(f(x')), e_j \rangle_{\mathcal{H}_k}^2 \right)^{1/2} \\
&= \| h(f(x)) \|_{\mathcal{H}_k} \| h(f(x')) \|_{\mathcal{H}_k} \\
&\leq \sup_{x \in \mathbb{X}} k(f(x), f(x)) < +\infty.
\end{aligned}
$$

另一方面，显然，

$$\begin{aligned}\langle H^\dagger A_{\text{neg}} H e_j, e_j\rangle_{\mathcal{H}_k} &= \langle A_{\text{neg}} H e_j, H e_j\rangle_{L^2(\mathbb{X},P_{\mathbb{X}})} \\ &= \int w(x)\langle h(f(x)), e_j\rangle_{\mathcal{H}_k} \int w(x')\langle h(f(x')), e_j\rangle_{\mathcal{H}_k} d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x) \\ &= \int\int \langle h(f(x)), e_j\rangle_{\mathcal{H}_k}\langle h(f(x')), e_j\rangle_{\mathcal{H}_k} w(x)w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x').\end{aligned}$$

因此我们得到，

$$\begin{aligned}\operatorname{Tr}(H^\dagger A_{\text{neg}} H) &= \sum_{j=1}^{\infty} \langle H^\dagger A_{\text{neg}} H e_j, e_j\rangle_{\mathcal{H}_k} \\ &= \int\int \langle h(f(x)), h(f(x'))\rangle_{\mathcal{H}_k} w(x)w(x') d\nu_{\mathbb{X}}(x) d\nu_{\mathbb{X}}(x') \\ &= \mathbb{E}_{x,x^-}\left[k(f(x), f(x^-))\right].\end{aligned}$$

因此，我们得到了所需结果，证明完毕。 □

G.2.1 从图割角度与相关工作的比较

HaoChen 等人(2021年) 已经研究了总体层面的谱聚类与对比学习之间的联系。然而，我们在第2节中引入的积分核（4.1 ）与 HaoChen 等人(2021年) 的积分核略有不同，因为：1）在第一项中，我们将 $w(x, x')$ 除以 ~~$w(x)w(x')$~~，而不是除以 $\sqrt{w(x)w(x')}$（参见 HaoChen 等人(2021年) 的附录 F）；2）我们还纳入了超参数 λ。

请注意，Tian (2022年) 提出了一个称为 α-CL 的统一框架，该框架从逐坐标优化的角度将各种对比损失联系起来。在 Tian (2022年) 中，对比协方差在理论分析中发挥着核心作用。另一方面，我们使用第 4 节中定义的相似度函数，因此我们的分析方法不同于 Tian (2022年) 中的对比协方差。

G.3 （39）的证明

（39） 的证明。为证明（39），我们紧密遵循Von Luxburg(2007年年第 5 节所介绍的方法。由于我们考虑的是总体层面的归一化割，为完整起见，我们给出（39）的证明。

定义恒等算子 $D: L^2(\mathbb{X}, P_{\mathbb{X}}) \to L^2(\mathbb{X}, P_{\mathbb{X}})$ 为 $D\psi = \psi$ for $\psi \in L^2(\mathbb{X}, P_{\mathbb{X}})$。根据 D 和 U的定义，我们有

$$DUe_i = \begin{cases} \frac{\mathbb{1}_{\mathbb{V}_i}(\cdot)}{\sqrt{\operatorname{vol}(\mathbb{V}_i)}} & (i \le K), \\ 0 & (i > K). \end{cases}$$

因此，对于 $i \le K$，我们有：

$$\langle U^\dagger DUe_i, e_i\rangle_{\mathcal{H}_k} = \int \frac{w(x)\mathbb{1}_{\mathbb{V}_i}(x)^2}{\operatorname{vol}(\mathbb{V}_i)} d\nu_{\mathbb{X}}(x) = 1.$$

另一方面，对于 $i \leq K$，我们有

$$\begin{aligned}&\langle U^\dagger A U e_i, e_i\rangle_{\mathcal{H}_k}\\ &= \int\int (w(x,x') - \lambda w(x)w(x')) \frac{\mathbb{1}_{\mathbb{V}_i}(x)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} \frac{\mathbb{1}_{\mathbb{V}_i}(x')}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x)\end{aligned}$$

因此，我们得到如下结果：

$$\begin{aligned}\mathrm{Tr}(U^\dagger(D-A)U) &= \sum_{i=1}^{\infty} \langle U^\dagger(D-A)Ue_i, e_i\rangle_{\mathcal{H}_k}\\ &= \sum_{i=1}^{K} \langle U^\dagger(D-A)Ue_i, e_i\rangle_{\mathcal{H}_k}\\ &= \lambda K + \frac{1}{2}\sum_{i=1}^{K} \int\int \big(w(x,x') - \lambda w(x)w(x')\big) \left(\frac{\mathbb{1}_{\mathbb{V}_i}(x)}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}} - \frac{\mathbb{1}_{\mathbb{V}_i}(x')}{\sqrt{\mathrm{vol}(\mathbb{V}_i)}}\right)^2 d\nu_{\mathbb{X}}^{\otimes 2}(x,x')\\ &= \lambda K + \sum_{i=1}^{K} \int_{x\in\mathbb{V}_i}\int_{x'\in\mathbb{V}_i^c} \frac{(w(x,x') - \lambda w(x)w(x'))}{\mathrm{vol}(\mathbb{V}_i)} d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x)\\ &= \lambda K + \sum_{i=1}^{K} \int_{x\in\mathbb{V}_i}\int_{x'\in\mathbb{V}_i^c} \frac{\mathrm{sim}(x,x';\lambda)}{\mathrm{vol}(\mathbb{V}_i)} w(x)w(x') d\nu_{\mathbb{X}}(x') d\nu_{\mathbb{X}}(x)\\ &= \lambda K + \sum_{i=1}^{K} \frac{W(\mathbb{V}_i, \mathbb{V}_i^c)}{\mathrm{vol}(\mathbb{V}_i)}.\end{aligned}$$

至此，证明完毕。 □

H 实验

H.1 实验设置

我们介绍本文实验的设置。实验所用代码基于 SimSiam[1] 的官方实现，并使用 PyTorch 编写（Paszke 等人，2019年）。我们基本遵循 Chen 和 He（2021）的实验设置。为确保内容完整，我们详细介绍实验中采用的设置。在预训练阶段，我们按如下方式构建可训练的编码器模型：遵循 Chen 和 He（2021），使用一个参数经过初始化的骨干网络架构，后接由线性层、批归一化（Ioffe 和 Szegedy，2015）和 ReLU 激活函数组成的多层感知机。请注意，这种多层感知机称为投影头（Chen 等，2020a）。与多项研究的做法一致，我们使用欧几里得范数对可训练编码器模型的输出进行归一化（Chen 等，2020a；Dwibedi 等人，2021年）。另一方面，在线性评估阶段（Chen 等，2020a；Chen 和 He，2021），我们从训练好的编码器模型中移除额外的多层感知机，然后向该编码器模型添加一个线性分类头。训练好的编码器模型的参数

[1] https://github.com/facebookresearch/simsiam (最后访问日期：2023年3月25日)

在此阶段保持冻结，只有线性头会接受训练。在本文报告的所有实验中，我们均使用 ResNet-18（He 等人，2016）作为骨干网络架构。投影头采用两层多层感知机，其中第一层线性层不带偏置，最后一层线性层带有偏置项。

在预训练中，我们采用与Chen 等人（2020b）相同的数据增强技术；Chen 和 He Chen 和 He（2021年）。请注意，遵循 Chen 和 He Chen 和 He （2021年）的做法，在 CIFAR-10 实验中，我们不使用高斯模糊增强。在线性评估中，我们也采用Chen 和 He （2021年）的数据增强技术。请注意，在预训练和线性评估两个阶段，我们都将训练数据加载器中的 drop last 设为 True。

在预训练和线性评估两个阶段的优化过程中，我们遵循Chen 和 He （2021年）的做法，使用 SGD 优化器。受 HaoChen 等人（ 2021 年）启发，我们使用带预热的余弦衰减学习率调度器（Loshchilov 和 Hutter，2017年）（Goyal 等人，2017年）。请注意，预训练和线性评估中我们都使用余弦衰减学习率调度器，并在预训练中应用预热。遵循 SCL 官方实现中学习率调度器的实现方式，[2]，我们还根据迭代次数定义学习率调度器。

H.1.1 配置

在本文报告的所有实验中，我们采用以下配置：

预训练 对于学习率，我们将初始学习率设为 0.0005，基础学习率设为 0.05，预热轮数设为 10。遵循 Chen 和 He (2021年），我们对学习率采用线性缩放（Goyal 等人，2017年）。对于 SGD 优化器的设置，我们同样遵循 Chen 和 He (2021年）在其 CIFAR-10 实验中采用的设置（见其论文附录 D）：动量设为 0.9，权重衰减设为 0.0005。对于编码器的输出维度，我们设为 512。

线性评估 我们遵循 Chen 和 He (2021年）在线性评估中的配置：对于 SGD 优化器，动量为 0.9，权重衰减为 0，批次大小固定为 256，学习率为 30.0，其中对学习率采用了线性缩放（Goyal 等人，2017年）。请注意，我们训练线性头 100 个训练轮数。

H.1.2 核函数

在实验中，我们使用以下核函数：

[2] https://github.com/jhaochenz/spectral_对比_学习/blob/ee431bdba9bb62ad00a7e55792213ee37712784c/ optimizers/lr_scheduler.py (最后访问时间：2023年3月25日)

高斯核。 高斯核 k_{Gauss} 定义为，

$$k_{\text{Gauss}}(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{\sigma^2}\right),$$

其中， $\sigma^2 > 0$ 是带宽参数。

二次核。 二次核 k_{Gauss} 定义为：

$$k_{\text{Quad}}(z, z') = \left(z^\top z'\right)^2.$$

H.1.3 损失函数

对于核对比损失的实现，我们实现了经验核对比损失（4）。请注意，在我们的实验中，使用高斯核和二次核的KCL框架分别称为高斯核对比学习（GKCL）和二次核对比学习（QKCL）。

为进行比较，我们还对SimCLR（Chen 等人，2020a）和SCL（HaoChen 等人，2021年）进行了若干复现实验。对于SimCLR目标函数的实现，我们使用Lightly的lightly.loss.NTXentLoss[3]（Susmelj 等人，2020年）。对于SCL谱对比损失的实现，我们采用了官方SCL代码的实现[4]。

H.1.4 数据集

在实验中，我们使用以下数据集：CIFAR-10（Krizhevsky，2009）、STL-10（Coates 等人，2011）和 ImageNet-100（Tian 等人，2020年）。需要注意的是，ImageNet-100 是 ImageNet-1K 数据集（Deng 等人，2009）的一个子集，其中包含归入 100 个类别的图像（Tian 等人，2020年）。从原始 ImageNet-1K 数据集中提取图像以创建 ImageNet-100 数据集时，我们选择了 Tian 等人（2020年）使用的 100 个类别。还需要说明的是，对于使用 STL-10 数据集的实验，我们使用由无标签图像和有标签训练图像组成的混合数据集进行预训练；在线性评估阶段，使用有标签训练图像训练线性头，并使用有标签测试图像计算线性评估的准确率。在所有实验中，我们对每个数据集使用以下图像尺寸： 32×32 像素（CIFAR-10）， 96×96 像素（STL-10），以及 224×224 像素（ImageNet-100）。CIFAR-10 和 STL-10 的图像尺寸分别与其原始图像尺寸相同；ImageNet-100 的图像尺寸则参考了 Chen 等人（2020a）；Chen 和 He（2021年）。

H.1.5 CIFAR-10 实验的架构详情

在 CIFAR-10 数据集上的实验中，我们遵循 Chen 和 He (2021年);HaoChen 等人 (2021年); He 等人 (2016) 的设置，对原始的 ResNet-18 (He 等人, 2016) 作如下修改：

[3]请注意，在我们的实验中，我们使用 Lightly 1.2.25。[4]https://github.com/jhaochenz/spectral_对比式_学习/blob/ee431bdba9bb62ad00a7e55792213ee37712784c/ models/spectral.py （最后访问时间：2023年3月25日）

在 TorchVision 的 ResNet 实现代码中[5] (TorchVision 维护者和贡献者,2016)，我们将第一个卷积层替换为卷积核大小为 3、步幅为 1、填充为 1 的卷积层，并将最大池化层替换为卷积核大小为 1、步幅为 1 的最大池化层。

H.1.6 实现的补充信息

我们在实验中使用以下软件包：PyTorch（Paszke 等人，2019年）、torchvi- sion（TorchVision 维护者和贡献者，2016）、NumPy（Harris 等人，2020年）、Lightly（Sus- melj 等人，2020年）、Matplotlib（Hunter，2007年）和 seaborn（Waskom，2021年）。

H.2 线性评估结果

我们使用 CIFAR-10、STL-10 和 ImageNet-100 数据集进行预训练和线性评估。在预训练阶段，我们对编码器模型进行 800 轮训练。对于 CIFAR-10 和 STL-10 数据集的实验，批次大小设为 256。此外，对于 ImageNet-100 数据集的实验，我们将批次大小设为 512。对于本小节报告的所有实验，我们选择以下 KCL 框架超参数： $\sigma^2 = 1$和 $\lambda = 8$ 分别用于 GKCL 和 $\lambda = 4$ 用于 QKCL。对于 SCL，我们受到 HaoChen 等人（2021年）的启发，将半径参数设为 3。对于 SimCLR，我们受到 Chen 等人（2020a）的启发，将温度参数设为 0。1。本小节的所有实验也都使用这些超参数。

结果见表 1。表 1中，CIFAR-10 数据集的实验使用一块 Quadro P6000 GPU 完成。此外，STL-10 和 ImageNet-100 数据集的实验使用一块 Tesla V100S GPU 完成。

H.3 更多实验

H.3.1 权重参数和批量大小的消融研究

我们研究 λ 和批量大小的选择如何影响学习到的表征质量。在实验中，我们使用 CIFAR-10 数据集。我们为 λ 选择 $\{1, 2, 4, 8, 16, 32\}$，并为批量大小选择 $\{64, 128, 256, 512, 1024\}$。每次运行中，我们对一个编码器模型进行 200个训练轮次的预训练。我们使用 GKCL 和 QKCL，并对结果进行评估。

[5] https://github.com/pytorch/vision/blob/eac3dc7bab436725b0ba65e556d3a6ffd43c24e1/torchvision/models/resnet.py (最后访问日期：2023年3月26日)

表1：各方法在线性评估中的 Top-1 和 Top-5 准确率（%）。对于 CIFAR-10 和 STL-10 实验，我们进行了三次“预训练+线性评估”试验，结果以均值±标准差表示。对于 ImageNet-100 实验，我们进行了一次“预训练+线性评估”试验。以下报告的所有结果均按如下方式获得：我们对线性头进行 100 个训练轮次的训练，并使用相应的验证数据集或测试数据集评估模型的最终分类准确率。“repro.”是“复现”的缩写，表示我们进行了多次复现实验，以与 KCL 的性能进行比较。

	CIFAR-10		STL-10		ImageNet-100	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
SimCLR（复现）	90.09±0.05	99.70±0.01	87.23±0.35	99.56±0.04	77.26	94.06
SCL（复现）	91.53±0.10	99.71±0.05	86.68±0.12	99.49±0.07	75.22	93.36
GKCL	90.87±0.08	99.63±0.00	86.69±0.09	99.38±0.01	76.40	93.20
QKCL	90.62±0.10	99.59±0.05	87.07±0.20	99.37±0.03	77.12	93.96

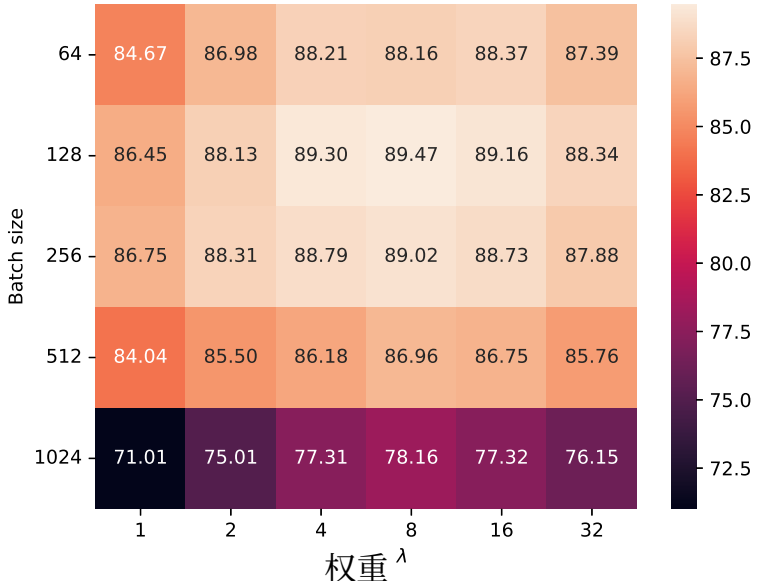


图 2：GKCL 的消融研究结果。每个单元格中的数字表示 Top-1 准确率（%）。

GKCL 和 QKCL 在线性评估结束时计算的 Top-1 准确率如图2和3所示。请注意，图2和3中的实验均使用一块 Tesla V100S GPU 完成。实验结果表明：1） λ 的取值会影响 KCL 学得的表示质量；2）预训练编码器时，较小的批量大小（例如 128 和 256）效率更高，而较大的批量大小（例如 1024）

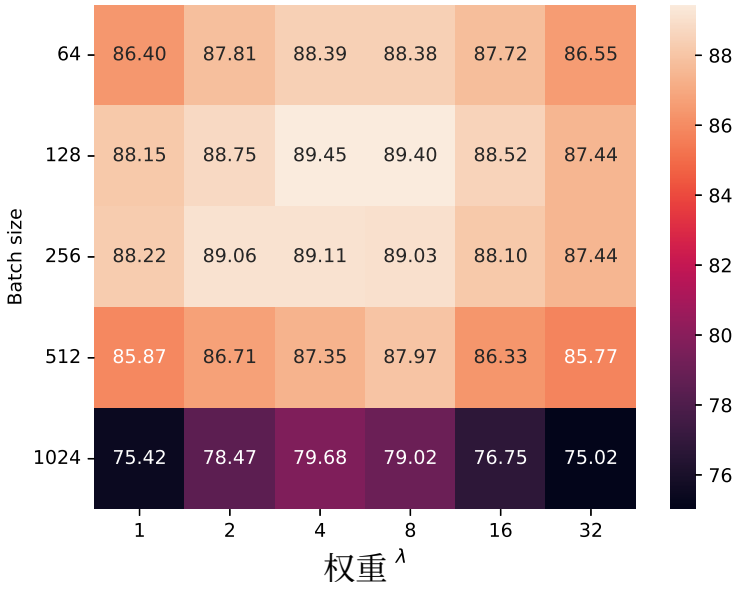


图3：QKCL 的消融研究结果。每个单元格中的数字表示 Top-1 准确率（%）。

会降低性能。请注意，Chen 等人 (2021年) 关于广义NT-Xent损失得出了与第一点类似的发现。此外，Chen 和 He (2021年) 指出，SimSiam 在小批量大小下效率更高。

H.3.2 λ 如何影响所学习表示的几何结构?

根据定理1，最小化核对比损失会使 $\lambda \cdot \mathfrak{c}(f$变小，这意味着随着 λ 增大，各簇的均值趋于均匀分布。受此结果启发，在本小节中，我们模拟属于各簇的特征向量均值的分布情况。在实验中，我们使用 STL-10 数据集。在无监督预训练阶段，我们使用 STL-10 数据集中无标签图像与有标签训练图像的组合。我们使用 GKCL 进行预训练。实验中使用的权重为$\{1, 2, 4, 8, 16, 32, 64\}$。每次运行中，我们对编码器模型进行 400 轮训练。批量大小设为 256。预训练阶段结束后，我们计算每个类别的均值，并计算这些均值之间的余弦相似度。由于 STL-10 数据集中的簇 $\mathbb{M}_1$，$\cdots$，$\mathbb{M}_K$ 难以获得，因此我们改用 STL-10 数据集中有标签训练图像所包含的标签，计算每个类别中由原始数据转换而来的增强数据特征向量的均值。请注意，计算均值时，我们从每张原始图像生成一张增强图像。本小节中的实验均使用一块 Tesla V100S GPU 完成。

结果汇总于表4 中，并以箱线图呈现。结果表明，随着 λ 增大，变化幅度会减小。因此，在这一实验设置下，较大的 λ 会使均值分布得更加均匀。请注意，这一结果可能意味着簇 $\mathbb{M}_1$，$\cdots$，$\mathbb{M}_K$ 以及

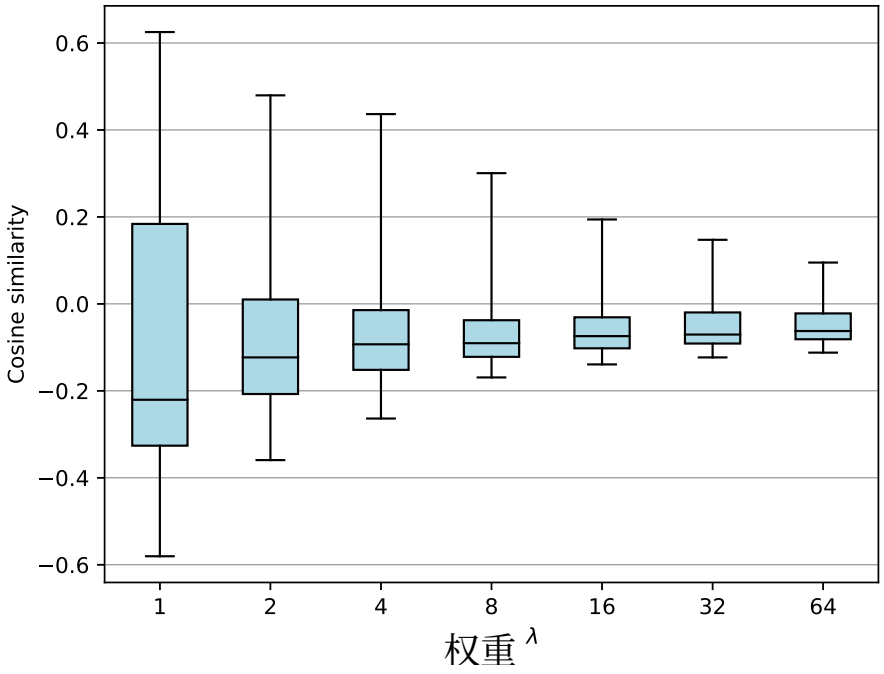


图 4：采用不同 λ 预训练的各编码器模型中，不同类别均值之间余弦相似度的箱线图。请注意，每个箱体中的水平线从下到上依次表示最小值、第一四分位数、中位数、第三四分位数和最大值。

由标签定义的子集之间存在某种关系。我们将这一问题的探究留作未来工作。

H.4 过大 的 λ 会降低下游分类任务的性能 s

在本小节中，我们报告了 λ 取不同值时的实验结果。实验中，我们将 $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ 用作权重 λ。对于对比学习框架，我们采用 GKCL。我们使用两个数据集 CIFAR-10 和 STL-10，并在每次运行中对编码器进行 400 轮预训练。在预训练阶段，我们将批次大小设为 128。本小节报告的每项实验均使用一块 Tesla V100S GPU 完成。

表2：线性评估结果中的 Top-1 准确率（%），其中编码器在每次运行中使用不同的 λ 进行预训练。

	Top-1 准确率	
λ	CIFAR-10	STL-10
1	89.06	83.20
2	90.29	84.54
4	90.77	85.36
8	90.66	85.33
16	90.52	84.19
32	89.78	83.39
64	88.85	81.50
128	86.93	79.71
256	85.24	77.89
512	82.43	76.26

线性评估结束时的 Top-1 准确率结果见表 2。结果表明，过大的 λ（例如 512）会降低下游任务的性能。