西安交通大學

博士学位论文

基于信息论的随机学习算法泛化理论研究

学位申请人：董裕欣

指导教师：李辰教授

学科名称：计算机科学与技术

2024 年 09 月

Information-theoretic Generalization Analysis for Randomized Learning Algorithms

A dissertation submitted to

Xi'an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

By

Yuxin Dong

Supervisor: Prof. Chen Li

Computer Science and Technology

September 2024

博士学位论文答辩委员会

基于信息论的随机学习算法泛化理论研究

答辩人：董裕欣

答辩委员会委员：

陕西师范大学教授：李永明 （注：主席）

西北大学教授：熊裕焱

西安交通大学教授：唐亚哲

西安交通大学教授：王嘉寅

西安交通大学教授：罗敏楠

答辩时间：2024 年 08 月 19 日

答辩地点：西安交通大学彭康楼 218 会议室

摘 要

近年来，以深度神经网络为代表的机器学习模型在金融、教育、生物等领域取得了巨大成功，但其背后的原理人们却知之甚少。随着相关技术的快速发展，人们迫切需要对其基本原理、能力与局限性进行深入理解。作为机器学习的基石，统计学习理论旨在通过理论分析系统地描述学习算法的行为。其中，一个关键问题是如何精准刻画随机学习算法的泛化能力。经典统计学习理论基于假设空间的相关复杂性度量，构建了一致收敛的泛化误差上界。算法稳定性分析方法则基于学习算法对训练数据扰动的敏感程度刻画其泛化能力。然而，这些方法均依赖于模型空间复杂度或特定强假设条件，在面对现代深度神经网络时往往无法适用或导向过于悲观的估计结果。

近期研究发现，利用训练数据与模型假设之间的互信息，能够在较弱的正则假设条件下基于数据分布与学习算法的特定属性构建泛化上界。进一步地，结合超样本技术下的条件互信息分析框架，可解决由连续可逆前向传播函数引起的无界互信息问题，得到基于网络预测值的可计算泛化上界。信息论相关泛化分析方法已获得国内外学者的广泛关注，成为分析现代机器学习模型泛化能力的主要途径之一。然而，现有理论结果在其泛化度量的可计算性、泛化估计的紧致性以及泛化结果的适用性等方面仍存在多种局限性。本论文聚焦于信息论视角下随机学习算法的泛化理论研究，深入探索多种经典学习场景下的泛化属性，突破当前理论结果的主要局限性，构建上界可计算、估计精度高、适用范围广的信息论泛化分析框架。本论文的具体研究内容可划分为：

（1）基于核化 Rényi 熵的可计算泛化误差估计。针对现有基于传统 Shannon 信息度量的泛化误差上界在实际应用中难以量化计算的问题，提出了一种新型信息度量准则——核化 Rényi 熵。其不受随机变量的维度所影响而能够直接通过有限采样数据点近似计算，并且兼容于现有基于 Shannon 熵的泛化分析框架。在此基础上，推广了现有面向 SGLD 与 SGD 等随机迭代学习算法的泛化理论，通过引入梯度轨迹协方差度量构建了更紧且可计算的泛化误差上界估计。进一步地，针对矩阵 Rényi 熵朴素算法计算效率低下的问题，基于矩阵迹估计与多项式近似技术构建了理论最优收敛阶的快速近似算法，为相关泛化上界的可计算性提供了切实保障。

（2）损失熵引导的高概率信息论泛化误差上界。针对现有基于高维互信息度量的泛化误差上界难以量化计算、估计不紧致的问题，提出了一种新型低维信息论泛化度量——损失熵。由于其仅包含一维随机变量，因此可通过核密度估计、分箱等方法直接量化计算，构建了一系列可计算的信息论泛化上界。在此基础上，将相关理论结果拓展至数据无关泛化场景，显著提升了基于信息瓶颈度量的泛化上界的紧致性与可计算性，

并为最小化误差熵准则的泛化能力提供了理论保证。进一步地，在数据依赖泛化场景中，收紧了现有留一法与超样本泛化分析框架下的泛化理论结果，构建了基于损失熵的高概率泛化误差上界。这是首个可计算的高概率信息论泛化上界，且在众多经典学习场景中相较于现有上界估计显著更紧。

（3）面向多点损失的一致信息论泛化分析框架。针对现有面向传统有监督单点学习的信息论泛化上界不再适用于对比学习等多点学习场景的问题，提出了首个信息论视角下的多点学习泛化理论框架。首先，通过互信息的超可加性进行期望泛化误差的拆分以及上界归约，克服了多点学习中的非独立同分布损失挑战。其次，通过对超样本变量的独立性拆分，构建了基于低维互信息度量的泛化误差上界，解决了在超样本框架下推广现有上界所面临的维度爆炸问题。这些结果适用于任意有界多点损失函数，能够将单点、双点、三点及更高阶情形学习范式纳入统一的泛化分析框架中。

（4）基于信息论的领域泛化理论与算法设计。针对现有基于传统独立同分布假设的泛化理论不再适用于领域泛化等分布外泛化场景的问题，构建了信息论视角下的领域泛化分析框架。首先，通过全局总体风险作为桥梁，将整体领域泛化误差分解为源域与目标域上的泛化误差，分别为其构建了信息论视角下的泛化上界，发现了影响学习算法领域泛化性能的关键互信息度量。随后，发掘了最小化相关泛化误差上界的关键因素，证明了梯度与特征对齐共同构成领域泛化的一种充分条件。基于此，研发了基于域间分布对齐的新型领域泛化算法，并改进了现有基于低阶矩的分布对齐算法，在 DomainBed 基准评估数据集上相较于朴素 ERM 方法将综合准确率由 68.7% 提升至 71.2%。

关 键 词： 信息论；统计学习理论；泛化分析；对比学习；领域泛化

论文类型： 理论研究

ABSTRACT

In recent years, machine learning models, especially deep neural networks, have achieved astonishing success in various fields, e.g. finance, education, and biology. Along with the development of these technologies, there is an increasing need for a deeper understanding of their underlying mechanisms, capabilities, and limitations. A key question in statistical machine learning theory is how to accurately characterize the generalization ability of randomized learning algorithms. Classical learning theory derives uniform convergence generalization bounds based on metrics of hypothesis space complexity. An alternative approach assesses generalization ability from the perspective of stability, which is evaluated as the sensitivity of the algorithm to certain perturbations in the training data. However, these methods often depend on the complexity of the hypothesis space or certain strong assumptions, making them unsuitable or overly pessimistic for modern deep neural networks.

Recent research has discovered that leveraging the mutual information between the training dataset and the hypothesis leads to generalization bounds under weaker assumptions, which can also characterize certain properties of the data distribution or the learning algorithm. Furthermore, by incorporating conditional mutual information analysis within the supersample framework, the issue of unbounded mutual information caused by continuous invertible forward propagation functions can be addressed, resulting in computationally tractable generalization bounds based on network predictions. Information-theoretic generalization analysis methods have gained significant attention among global scholars, and have become one of the primary approaches for analyzing the generalization abilities of modern deep learning models. However, existing works still have various limitations in terms of computational tractability, tightness of generalization bounds, and applicability of theoretical results. This dissertation focuses on information-theoretic generalization analysis for randomized learning algorithms, exploring generalization properties in various classical learning scenarios, overcoming major limitations of current theoretical results, and constructing a computable, high-precision, and broadly applicable information-theoretic generalization analysis framework. The major contributions of this dissertation can be divided into:

(1) Computationally Tractable Generalization Bounds via Kernelized Rényi's Entropy. To address the issue that existing generalization bounds based on traditional Shannon's in-

formation measures are computationally intractable in practice, a novel information measure, named kernelized Rényi's entropy, is proposed. It can be directly approximated with finite data points, regardless of the dimensionality of the corresponding random variables, and is still compatible with the existing generalization analysis framework based on Shannon's entropy. The existing generalization results for iterative and noisy learning algorithms, e.g. SGLD and SGD, are successfully extended with kernelized Rényi's entropy, resulting in tighter and computationally tractable generalization bounds based on gradient covariance metrics. Additionally, to address the computational burden of eigenvalue decomposition required by matrix-based Rényi's entropy, fast approximations with theoretically optimal convergence rates based on trace estimation and polynomial approximation techniques are proposed, ensuring the computability of related generalization results.

(2) Loss Entropy Induced High-probability Generalization Bounds. To address the issue that existing generalization error bounds based on high-dimensional mutual information measures are computationally intractable and need to be tightened, a novel low-dimensional information-theoretic generalization measure, named loss entropy, is proposed. It only involves one-dimensional random variables, and thus can be directly estimated using kernel density estimation or binning methods, completely solving the problem of computational intractability in related works. Upon loss entropy, relative theoretical results are successfully extended to data-independent generalization scenarios, significantly improving the tightness and computational tractability of generalization bounds based on information bottleneck measures, and providing new theoretical insights into the generalization behavior of the minimum error entropy criterion. Furthermore, existing generalization results for data-dependent generalization scenarios are improved under the leave-one-out and supersample frameworks, providing high-probability generalization bounds based on loss entropy. This is the first computationally tractable high-probability information-theoretic generalization bound, and is also significantly tighter than existing works in various classical learning scenarios.

(3) A Unified Generalization Framework for Non-pointwise Learning. To address the issue that existing information-theoretic generalization bounds for traditional supervised pointwise learning scenarios are no longer applicable to non-pointwise learning scenarios, e.g. contrastive learning, the first information-theoretic generalization framework for non-pointwise learning paradigms is proposed. Firstly, by utilizing the superadditivity of mutual information to decompose the expected generalization error and adopting a bottom-to-top reduction, the challenge of

non-i.i.d. loss terms in non-pointwise learning is overcome. Next, by adopting an independent decomposition of the supersample variables, generalization bounds based on low-dimensional mutual information measures are derived, completely solving the dimensional explosion challenge in extending existing results within the supersample framework. The results apply to any bounded non-pointwise loss functions, encompassing pointwise, pairwise, triplet, and higher-order learning paradigms, all within a unified framework.

(4) Information-theoretic Analysis and Algorithm Design for Domain Generalization. To address the issue that existing generalization bounds based on the traditional i.i.d. sample assumption are no longer applicable to domain generalization or other out-of-distribution scenarios, a high-probability domain generalization analysis framework is proposed under the lens of information theory. Firstly, using the global population risk as a bridge, the domain generalization error is decomposed into source-domain and target-domain generalization errors. The following analysis then derives information-theoretic generalization bounds for them respectively, identifying the key mutual information measures that affect the domain generalization performance of learning algorithms. Subsequently, aligning inter-domain gradients and representations jointly constitutes a sufficient condition for domain generalization. Based on these findings, the IDM algorithm is proposed for domain generalizaion, and the PDM method improves over existing distribution matching methods based on low-order moments, together improving the overall performance from 68.7% to 71.2% compared to the traditional ERM method on the DomainBed benchmark.

KEY WORDS: Information Theory; Statistical Learning Theory; Generalization Analysis; Contrastive Learning; Domain Generalization

TYPE OF DISSERTATION: Theoretical Research

目 录

CONTENTS

1 绪论

1.1 研究背景

近年来，数据、算力、算法的快速迭代发展，促使了以深度神经网络 (Deep Neural Networks) 为代表的机器学习模型在金融、教育、生物等诸多应用领域的快速落地。人工智能和机器学习已经成为各个领域变革性进步的推动力量，在人们的日常生活和众多基础行业中逐渐普及开来。然而，除却机器学习技术已取得的巨大成功，其背后的深层原理人们却知之甚少。随着相关技术的不断发展，人们迫切需要对其基本原理、能力和局限性进行深入理解。2017 年，国务院在《新一代人工智能发展规划》中强调了“实现具备高可解释性、强泛化能力的人工智能”的目标。同年，图灵奖得主姚期智院士在接受采访时表示：“人工智能的下一个突破口在于理论”。2019 年，《人民日报》发文“根据人工智能发展规律，每隔十几年往往会出现引领人工智能整体发展的新因素，而当下新一代人工智能的红利释放时间有限，可能很快就会触碰到发展的瓶颈。因此，我们必须抢占下一代人工智能发展先机，在理论、方法、工具、系统等方面走在前面、占领制高点”。2022 年，鄂维南院士先后在国际数学家大会以及国际机器学习大会上做了关于《从数学视角看机器学习》和《迈向机器学习的数学理论》的报告，阐明了数学理论和机器学习发展的时代背景与共同主线。

作为机器学习发展的基石，统计学习理论旨在建立普适的机器学习理论分析框架，并系统地描述随机学习算法的行为。其中，学习算法是一类选择规则集，其基于所给定的训练数据集在假设空间中选择合适的假设。在统计学习理论中，一个关键问题是如何精准刻画随机学习算法的泛化能力 (Generalization Ability)。泛化分析的目标是为学习算法的泛化能力提供理论保证，即当所选择的假设面对训练数据集中未出现的新样本时，是否仍然能够保持良好的性能（通常通过损失函数衡量）。经典的统计学习理论基于近似可能正确 (Probably Approximately Correct, PAC)[1]的形式化可学习性框架，通过刻画数据分布、模型架构与假设空间之间的关联性，建立了基于假设空间复杂度的泛化上界。具体而言，证明一类假设的 PAC 可学习性可归结为一种强一致收敛结果，其要求对于任意数据分布，均存在一个学习算法，使得在提供足够训练数据的前提下，其性能能够无限趋近于该假设空间中的最优总体风险。事实证明，PAC 可学习性等价于一致收敛性 (Uniform Convergence)[2]。因此，若某一类假设满足一致收敛性，则能够推导出在数据分布和假设上均一致的泛化上界，使得对于任意给定的数据分布与学习算法，其均能够保证训练误差在训练数据充足时趋近于总体期望误差。上世纪中，Vapnik 等[3-4]研究了一类机器学习模型的一致收敛性，后被称为 Vapnik-Chervonenkis (VC) 维度，可看作是一类假设空间复杂度的度量。Blumer 等[5]进一步建立了两者间的理论联

系，证明了假设空间的 VC 维度能够表征其 PAC 可学习性。直观而言，VC 维度衡量了假设类在任意给定特征标签下所能够拟合的最大数据集大小。若样本总数大于此值，模型将无法拟合全部的数据样本，而剩余样本则可作为其期望损失的合理估计。由此，假设空间的 PAC 可学习性完全由其 VC 维度所决定。然而，一致收敛性对于多数现代深度学习模型而言过于严苛。此类模型通常能够良好拟合自然发生的数据分布，而仅在某些边界数据或假设条件下表现出较差的泛化能力。这启发了后续基于特定数据分布或学习算法的泛化度量，这些方法通常能够在一定程度上弱化其前提假设。

在经典统计学习理论中，另一个重要的泛化度量是 Rademacher 复杂度[6-7]。相较于 VC 维度，假设空间的 Rademacher 复杂度通常基于给定的数据集（或期望意义下的数据集分布）而定义。直观而言，通过将数据集随机划分为训练集和测试集，Rademacher 复杂度刻画了在该假设空间中的最坏情形下，训练集与测试集上损失之间的平均差异。Rademacher 复杂度不仅可用于推导一类假设空间（如支持向量机[2]）的泛化上界，且针对有限 VC 维度的假设空间，其能够导出更紧的泛化估计结果。然而，虽然 Rademacher 复杂度引入了数据分布依赖，其刻画的依然是某个假设空间内的最坏情形，在面对现代机器学习算法时通常会导向过于悲观的泛化估计结果。深度神经网络由于其复杂的内蕴结构，往往能够在拟合全部训练数据的同时，依旧表现出良好的泛化能力。这种表现被称为深度学习中的过参数化 (Over-parameterization) 或双下降 (Double Descent) 现象[8-9]，即在网络复杂度达到一定规模后，模型的训练误差与测试误差同时随其表达能力的进一步提升而下降。经典统计学习理论将泛化归因于假设空间复杂度的制约，从而无法对此类常见的过参数化现象做出有效解释。这是 VC-维度、Rademacher 复杂度与覆盖数 (Covering Number) 等经典泛化度量的主要缺陷所在。

与此同时，另一类泛化分析方法从学习算法的稳定性角度出发，基于算法本身对于训练数据扰动的敏感程度刻画学习算法的泛化能力。直观而言，若算法所选择的假设对具体训练数据不存在强依赖性，则应具备良好的泛化能力。对此类强依赖性的形式化定义包括一致稳定性 (Uniformly Stability)[10-11]、参数稳定性 (Argument Stability)[12]、平均稳定性 (On-average Stability)[13]等。如 Shalev 等[14]所示，算法稳定性与一致收敛性之间也存在根本性联系。对于多数具备稳定性的学习算法（如线性回归、支持向量机等），其稳定性参数随训练数据量的增加而衰减，使得对应的泛化误差上界随之趋近于零。这些泛化上界不仅可用作机器学习模型在新数据上性能表现的理论保障，更可促进对相关学习算法的深入理解，为新型算法的设计提供灵感，以求进一步改善其泛化能力。然而，基于稳定性的分析框架往往依赖于机器学习模型的特定性质，这种依赖将导致泛化上界难以适用于现代神经网络。具体而言，基于一致稳定性的泛化分析技术通常假设网络的损失景观 (Loss Landscape) 满足 Lipschitz 连续、光滑、（强）凸等前提条件，而其对应的条件常数在现代深度神经网络中往往不存在或难以估计。进一步

地，斯坦福大学著名机器学习专家 Benjamin Roy 指出[15]：若不借助参数计数 (Counting Parameters) 或样本复杂度关于网络深度呈指数级的上界，现有的理论分析框架难以对超过两层以上网络的泛化性进行解释。因此，亟需探索新型泛化理论分析技术。

本论文将重点关注泛化理论分析中基于信息论的分析技术，和与之强相关的 PAC-Bayesian 泛化分析技术。长期以来，广大学者对于泛化与信息间关联的探索从未间断：奥卡姆剃刀原理[16]认为，简洁的解决方案通常比复杂的方案具备更优秀的泛化能力。基于这种思想，国内外学者提出了各式各样的形式化复杂度度量，用以捕捉某种类型的“信息量”准则。最初的探索可追溯到 Edgeworth 等[17-18]的 Fisher 信息量，Shannon[19]的信息论，以及 Kolmogorov 等[20]的 Kolmogorov 复杂度。Yang 等[21-22]开创性地通过此类复杂度度量建立了概率密度估计的性能保障。在统计学习背景下，信息的概念还包含了 Akaike[23]的 Akaike 信息准则，Schwarz[24]的贝叶斯信息准则，以及 Rissanen 等[25]研究的最小描述长度准则。本论文重点关注的信息论泛化方法可追溯到 Zhang[26]的工作，以及近期 Russo 等[27]和 Xu 等[28]的开创性工作。根据其研究思想，学习算法将被视为从训练数据到假设的通信通道，而通信中应用的信息度量（如熵、互信息等）在分析中起到了关键作用。PAC-Bayesian 方法则由 McAllester[29]和 Shawe 等[30]开创，后由 Catoni 等[31]发展完善。基于信息论的泛化分析技术显式考虑了数据样本的概率分布，并结合假设分布以分析具体的学习算法。相较于经典复杂度理论或算法稳定性理论，信息论分析方法更适用于现代深度学习模型，具备更强的理论可解释性。

1.2 研究目的、挑战与意义

基于信息论的泛化分析方法获得了国内外学者的广泛关注，已成为分析现代深度学习模型泛化能力的主要途径之一。然而，现有工作在信息度量的可计算性、泛化误差估计的紧致性以及理论结果的适用性等方面仍存在多种局限性，具体表现如下：

（1） **泛化上界难以量化的局限性：**目前，基于信息论的随机学习算法泛化上界多由假设与训练样本间的 Shannon 互信息所界定[28,32]。然而，对于深度神经网络模型而言，其对应的互信息上界往往以高维形式呈现。由于 Shannon 熵定义的固有局限性，此类高维互信息度量往往难以量化计算，甚至会出现无界情形[33]，导致此类上界的理论值与实际计算结果存在显著差距，无法准确反映模型的实际泛化能力。

（2） **泛化上界估计不紧致的局限性：**现有面向具体学习算法的信息论泛化上界往往依赖于损失函数的 Subgaussian 条件以及一些关键度量，包括局部梯度敏感性 (Local Gradient Sensitivity)、局部梯度方差 (Local Gradient Variance)、局部损失敏感性 (Local Value Sensitivity) 等[34]。基于此类假设与度量的泛化上界往往严重偏离了真实泛化误差值，从而失去了实际指导意义。

（3） **理论结果适用范围的局限性：**一方面，当前基于信息论的泛化分析框架仅聚焦于经典的有监督单点学习 (Pointwise Learning) 场景，无法拓展到需要考虑样本间耦合关系的双点学习 (Pairwise Learning) 或更一般的多点学习情形，包括对比学习 (Contrastive Learning)、度量学习 (Metric Learning)、排序算法 (Ranking Algorithm) 等常见学习场景；另一方面，目前的信息论泛化理论结果多基于数据样本的独立同分布 (Independently and Identically Distributed, i.i.d.) 假设，在部分真实学习场景中难以得到满足。

针对泛化上界难以量化的局限性，本论文首先引入新型可计算熵度量准则，将其推广至无限样本情形，作为可量化泛化上界的构成基础，并结合矩阵迹估计以及多项式近似技术设计该度量的理论最优近似算法，提高相关上界的可计算性；进一步地，结合损失熵构建新型低维高概率泛化分析框架，在数据无关及数据依赖场景下分别推导可计算的泛化误差上界，同时为理解最小化误差熵 (Minumum Error Entropy) 准则提供了新的信息论视角。针对泛化上界估计不紧致的局限性，本论文进一步松弛了对损失函数的有界性约束，减轻模型对先验知识的依赖，同时引入基于信息论的典型子集/超样本集 (Supersample Setting) 分析技术，得到更紧致的泛化上界估计；进一步地，针对基于梯度的随机迭代优化算法，引入梯度协方差矩阵替代现有的梯度方差度量，以求更精确地刻画其学习轨迹 (Learning Trajectory)，阐明影响算法泛化性能的关键因素。针对理论结果适用范围的局限性，本论文面向领域泛化问题构建了信息论泛化分析框架，突破当前基于平均或最坏情形的理论分析定式，提出了基于分布对齐的领域泛化算法，实现域不变 (Domain Invariant) 特征的自动发现；进一步地，构建突破单点限制的信息论泛化分析框架，探索新型样本解耦与降维分析技术，拓展现有的信息论单点泛化结果至双点及任意高阶情形，为对比学习、度量学习等多点学习场景提供理论保障。

2020 年 IEEE James Massey 奖获得者，麻省理工学院教授 Yury Polyanski 在其专著《Information Theory: From Coding to Learning》中指出：信息论与统计机器学习相互交融发展已成为常态，自信息论诞生以来，其在理解以及突破统计机器学习的基本限制方面一直不可或缺，互信息 (Mutual Information)、f-散度 (f-Divergence)，度量熵 (Metric Entropy) 等信息度量准则被广泛应用于构建统计估计的最小最大收敛速度 (Minimax Rate)。通过发展新的信息论分析框架，能够为解释随机学习算法，尤其是深度神经网络的泛化能力提供新视角和新思路，并启发面向领域/分布外 (Out-of-Distribution, OOD) 泛化任务的算法设计及理论构建，具有重要科学价值。

理论上，本研究有助于建立信息论创新驱动的可量化泛化理论体系，加深对影响随机学习算法泛化性能关键因素的理解；有助于构建基于信息论的概率领域泛化理论，突破当前基于平均或最坏情形的领域泛化分析框架，推动面向信息论的机器学习基础理论与方法研究；有助于构建面向多点学习的信息论泛化分析框架，突破当前基于假

设空间复杂性以及稳定性的理论分析定式。算法上，本研究有助于设计具有统计理论保障的优化与近似算法，拓展现有随机数值线性代数理论工具，得到具备良好统计性质的估计结果；进一步地，启发设计高效、稳健的领域泛化学习算法，实现域不变特征的自动发现，提升随机学习算法的分布外泛化性能。综上所述，研究信息论视角下随机学习算法的泛化理论，能够在较弱的假设条件下得到更为精确的泛化误差上界，发掘影响泛化性能的关键因素，从而更好地分析与控制算法的泛化行为。对以上关键问题的突破，不仅能另辟蹊径，建立基于信息论刻画学习算法泛化性能的理论基础，也将为现代人工智能理论的发展提供新思路和新方法。

1.3 国内外研究现状

以基于信息论的泛化理论分析为主线，本节详细梳理了信息论相关的泛化分析理论结果与工具的发展主线，并依据其应用场景、技术手段、发展阶段等划分如下：

1.3.1 基于信息论的统计学习理论概述

信息论最初由 Shannon[19]在 20 世纪 40-50 年代建立，为信息的表示、处理、存储和传输提供了一个严格的数学框架。Shannon 通过多种不同的信息度量刻画了这些信息相关过程的极限所在：熵刻画了完美重建约束下信息的压缩极限，在分布不匹配时可改用相对熵（或 KL 散度，Kullback–Leibler Divergence）度量；互信息则刻画了信息在不可靠媒介上可靠传输的极限。

直观而言，学习算法的泛化能力体现在其捕捉不同训练样本间的共性，而忽略单个样本特异性的能力。这可视为奥卡姆剃刀原理的一种应用，即在所有能够良好拟合训练集的假设中，应当优先选择最简假设。对“最简”的一种理解是，其在训练数据中所提取的信息越少，则该假设越简洁。基于信息论的泛化分析技术通过信息论度量刻画（随机）学习算法的泛化误差，通过形式化定义的互信息验证了这一原理。与基于一致收敛的泛化分析不同，信息论泛化上界不仅衡量了相关假设空间的复杂度，还包含对特定学习算法与数据分布的依赖。这种特性催生了基于信息论以及 PAC-Bayesian 泛化分析技术的、目前最精确的神经网络泛化误差上界保证。

基于以上思想，Xu 等[28]基于有界变量的次高斯性和相对熵的 Donsker-Varadhan 变分表示[35]，通过假设和训练数据间的互信息刻画学习算法所提取的信息量，构建了首个基于 Shannon 互信息度量的泛化误差上界。该工作可视为 Russo 等[27]理论结果的一种一般性拓展，其主要结果是在测量值符合次高斯性 (Subgaussianity) 的假设下，基于互信息度量构建了自适应数据分析 (Adaptive Data Analysis) 中的测量值偏差上界。此类问题定义可视为统计学习的一种等价情形，其中测量值对应于损失，而分析索引对应于有限假设空间中的一个假设。值得一提的是，信息论相关泛化分析理论的发展在

此前很大程度上独立于 PAC-Bayesian 相关研究主线。Russo 等[27]首次注意到其与 PAC-Bayesian 泛化上界间的相似性，并指出 PAC-Bayesian 与自适应数据分析之间的实质联系。Xu 等[28]的工作建立了统计机器学习与 Russo 等[27]的结果之间的关联，并将其假设空间推广到了任意不可数假设类。在此之前，Raginsky 等[36]推导了信息论版本的算法稳定性泛化上界，其以假设和单个训练样本间的互信息为主要度量。此类方法的核心思想可追溯到 Shawe 等[30]的工作，其通过一种类似于先验的幸运度量推导贝叶斯预测器的 PAC 上界。McAllester[29]进一步将其推广到一般情形。后续研究在此基础上继续拓展[31,37-39]，进一步提升了其适用范围与紧致性。此前，PAC-Bayesian 方法多针对有界损失对泛化误差进行刻画，其中最常用的是 0-1 损失。Zhang[26]基于其信息指数不等式 (Information Exponential Inequality) 相关结果，发展了面向通用损失函数的泛化上界。

信息论方法中，另一个受到广泛关注的统计机器学习技术是信息瓶颈 (Information Bottleneck) 方法[40]。具体而言，考虑随机变量 X 和 Y，其中 X 是输入，Y 是输出。其目标在于寻找一个表示 T，它是 X 的压缩版本，且应有助于预测 Y。信息瓶颈的基本思想是寻找最优的条件分布 $P^*_{T|X}$，使得对于给定超参 $\beta > 0$，有

$$P^*_{T|X} = \arg\max_{P_{T|X}} \{I(T;Y) - \beta \cdot I(X;T)\}. \tag{1-1}$$

其中，互信息 $I(T;Y)$ 衡量了表示 T 的充分性 (Sufficiency)，即其所保留的有助于预测 Y 的信息量；$I(X;T)$ 则衡量了 T 的最小性 (Minimality)，即其所继承的原始输入 X 中的信息量。参数 β 控制了两者之间的平衡。由信息论中的压缩机制驱动，Shwartz 等[41]认为信息瓶颈也有助于解释统计学习，特别是神经网络中的泛化现象。经过实证研究，Shwartz 等[41]认为神经网络的训练过程由两个阶段组成：首先是拟合阶段，此阶段中 $I(T;Y)$ 与 $I(X;T)$ 均在增加，使网络获得良好的预测性能；之后是压缩阶段，此阶段中 $I(T;Y)$ 不变但 $I(X;T)$ 开始减小，使得网络学习到压缩的、泛化良好的特征表示。Achille 等[42]进一步发展了这一理论，通过额外的正则化项获得良好压缩的特征表示，并建立了其与 PAC-Bayesian 理论的联系。Saxe 等[43]则质疑了拟合与压缩阶段的存在，认为此类现象并非普遍存在，且在很大程度上取决于具体实现细节。Goldfeld 等在后续工作[44-45]中进一步发展了基于信息瓶颈的机器学习方法。Kawaguchi 等[46]建立了基于信息瓶颈的泛化理论上界。

1.3.2 期望意义下的信息论泛化误差上界

在信息论泛化研究文献中，广泛考虑的一类目标是建立期望意义下的泛化误差上界。Russo 等[27]及 Xu 等[28]的工作均聚焦于此种情形，使训练数据与假设间的互信息成为期望泛化误差的一种基本度量方法。其基础数学工具是基于 Donsker-Varadhan 变分表示对相对熵的一种重新表述，这种思想也可见于众多面向通用函数的 PAC-Bayesian

泛化上界[47-49]。Rodríguez 等[50]在后续工作中探索了多种不同形式的次高斯性假设，这些假设均可导向与 Xu 等[28]的工作相似的泛化上界。Hellström 等[51]推导了基于二值 KL 散度的泛化上界，这种方法可追溯到 McAllester[52]的工作，其能够在训练集拟合较好时给出更精确的泛化误差估计。Jiao 等[53]则首次基于 f-散度给出了自适应数据分析的泛化上界，其可进一步推广至一般的统计学习情形。

然而，基于训练数据与假设间互信息的泛化上界在特定情形下可能出现无界情形。例如当训练数据与假设均为连续型随机变量，且其映射关系为确定型映射时，Donsker-Varadhan 变分表示所要求的概率密度绝对连续性 (Absolute Continuity) 将不再成立。Bu 等[54]探索了一种基于随机子集 (Randomized-subset) 技术的解决方案，利用期望算子的线性性质推导训练集随机子集损失的期望上界。这种方法的一种极限情况称为个体样本 (Individual-sample) 方法，即随机子集均为随机选择的单个样本。此类技术随后由后续工作[33,50,55-57]推广至其他应用场景中。Harutyunyan 等[58-59]继而探索了随机子集方法的数个关键性质，指出了随机子集大小对泛化上界估计的影响，以及个体样本互信息无法用于推导平方泛化误差的期望上界。Aminian 等[60]转而从概率度量本身考虑随机子集，在特定场景下导出了更紧的上界。

Raginsky 等[36]进一步基于 Wasserstein 距离定义了算法稳定性，首次推导了基于 Wasserstein 距离的泛化上界。具体而言，若训练集中的某个样本被替换为其他样本，学习算法所选择的假设分布不应发生较大变化（通过 Wasserstein 距离度量）。Wintenberger[61]基于弱传输不等式推导了具有快速收敛率的 Oracle 不等式。Lopez 等[62]和 Wang 等[63]各自独立导出了基于训练数据与假设的条件分布和边缘分布间 Wasserstein 距离的泛化结果，这些结果后续在 Rodríguez 等[64]的工作中与个体样本技术相结合，得到了更精确的泛化误差估计结果。Aminian 等[60]进一步利用概率度量的凸性，为非对称学习算法推导了更优的泛化上界。最后，Clerico 等[65]将链式技术推广到互信息以外的信息度量，证明了基于 Wasserstein 距离与 f-散度的链式泛化上界。

其他面向期望意义下泛化误差刻画的理论工作还包括：Alabdulmohsin[66]提出了一种面向所有参数化损失函数的一致泛化概念，并展示了其类似于 Xu 等[28]工作基于全变差 (Total Variation) 的一种变体。Hafez-Kolahi 等[67]从图模型角度讨论了基于条件与处理技术进一步收紧信息理论泛化上界的方法。Aminian 等[68]提出了以 Jensen-Shannon 散度为基础的上界，其可看作是相对熵的一种对称性拓展。Modak 等[69]基于 Rényi 散度推导了 Xu 等[28]结果的变体，在某些特定情形下可导向更紧的泛化上界。Aminian 等[70]考虑了泛化误差的高阶矩上界，提供了基于诸如互信息和 χ^2-散度等多种信息度量的上界。Raginsky 等[71]对基于信息论稳定性的泛化上界进行了全面讨论。Sefidgaran 等[72]引入了率失真 (Rate-distortion) 相关理论工具。Esposito 等[73]的结果为泛化上界与运输成本不等式 (Transportation-cost Inequalities) 的推导提供了新视角，该框架可推广至基于任意

散度度量（如互信息）的泛化上界。Wongso 等[74-75]考虑了基于一维随机投影的切片互信息，并从理论角度建立了其与泛化之间的联系。Chu 等[76]通过测度变换和 Young 不等式提供了一种推导信息论泛化上界的一致框架。Hafez 等[77]和 Xu 等[78]推导了贝叶斯学习中最小超额风险 (Minimum Excess Risk) 的上界。后续 Hafez 等[79]推导了最大最小超额风险的信息论上界，Dogan 等[80]则推导了期望超额风险的下界。

1.3.3 基于信息论的高概率泛化误差上界

期望意义下的泛化分析所提供的上界估计是在整体数据集与学习算法联合分布下对所有可能情形取期望的结果。在实际场景中，人们通常仅能够得到训练集的一个实例，而无法得到真实的数据分布。此类情况下，人们所关注的是学习算法在此特定数据集上的泛化表现，而非在整个数据分布上的平均表现。此外，由于计算资源的限制，人们通常仅关注单次运行学习算法所得到的特定假设的泛化性能，而非多次运行学习算法的平均性能。因此，另一种研究目标是在数据分布上以一定概率刻画泛化误差的理论上界。在以 Shawe 等[30]和 McAllester[29]为代表的 PAC-Bayesian 泛化分析文献中，多数泛化上界刻画的是对于任意给定的数据集，学习算法的条件期望泛化误差。另一种基于概率的泛化上界是在训练集与假设的联合分布下刻画泛化误差的概率分布，称为单次抽样 (Single-draw) 情形。此类上界在信息论和 PAC-Bayesian 文献中也偶有出现。

利用指数随机不等式 (Exponential Stochastic Inequality) 进行泛化上界推导的思想可追溯到 Zhang[26]和 Catoni[31]的开创性工作，这些上界在后续工作[81-84]中得到形式化。PAC-Bayesian 泛化上界的一般形式由 Germain 等[49,85-86]给出，后续工作通过结合有界损失函数的次高斯性[87]进一步给出了基于相对熵以及二值 KL 散度的泛化上界[37,88]。Foong 等[89]探讨了此类上界的紧致性及其对数依赖的进一步改进空间。McAllester[52]通过参数化二值 KL 散度推导了更紧的泛化上界。Hellström 等[90]进一步研究了泛化上界中通用凸函数的作用，并给出了理论最优选择。Jang 等[91]通过引入在线学习 (Online Learning) 中的投币 (Coin-betting) 框架改进了有界损失的 PAC-Bayesian 泛化上界。Bégin 等[47]将上界中的信息论度量从相对熵拓展至 Rényi 散度，并由 Alquier 等[48]进一步拓展至有界损失情形。Ohnishi 等[92]对基于 f-散度的测度交换不等式 (Change of Measure Inequality) 及其在 PAC-Bayesian 泛化上界中的应用进行了全面讨论。基于数据划分的数据依赖先验由 Ambroladze 等[93]引入，并在后续工作中得到多次拓展[49,82,94-97]。Seeger[98]使用了一种相似的技术，通过一组独立的模型选择样本以学习先验。基于差分隐私 (Differential Privacy) 的数据依赖先验由 Dziugaite 等[95]提出，而后 Rivasplata 等[49]提出了基于算法稳定性的先验。Catoni[31]与 Lever 等[99-100]则讨论了分布依赖的先验。

在主流的 PAC-Bayesian 泛化分析方法之外，Catoni[31]通过类似技术建立了基于后验单次抽样的泛化上界。Rivasplata 等[49]建立了单次抽样泛化的通用不等式，其在 Hell-

ström 等的后续工作[56,87,101]中得到进一步推广与讨论。Esposito 等[102]通过 Hölder 不等式进一步拓展了单次抽样的泛化结果。Xu 等[28]基于 Bassily 等的工作[103]将期望泛化上界转换为单次抽样上界。

其他面向高概率泛化误差上界刻画的理论工作还包括：Germain 等[85,104]讨论了 PAC-Bayesian 上界与贝叶斯推理和 KL 正则化的联系。Catoni 等[105]讨论了亚指数变量的 PAC-Bayesian 上界。Alquier[106-107]利用截断损失处理无界损失函数，Catoni 等[108]则使用了一种稳健损失函数以处理重尾分布。Holland[109]推导了重尾损失的 PAC-Bayesian 上界，并在此基础上建立了一种新的 Gibbs 后验。Biggs 等[110]通过问题的内在难度刻画得到了基于超额风险的更紧的上界。Herbrich 等[111]和 Langford 等[112]推导了基于学习预测器边距的上界，这种方法最近被 Biggs 等[113]用于建立去随机化的泛化上界。Audibert 等[114]将链式技术与 PAC-Bayesian 技术相结合。类似地，Asadi 等[115]基于多层相对熵推导了链式泛化上界，而 Clerico 等[65]则推导了另一种链式 PAC 上界。Yang 等[116]基于 Rademacher 过程推导了快速收敛率的 PAC-Bayesian 上界。Saunshi 等[117]推导了有关 Rademacher 复杂度的无监督对比表示学习 (Contrastive Unsupervised Representation Learning) 的泛化上界。其结果继而被 Nozawa 等[118]推广，得到了适用于非独立同分布数据的 PAC-Bayesian 泛化上界，并指导了新型表示学习算法设计。Mhammedi 等[119]推导了基于条件风险值 (Conditional Value at Risk) 的 PAC-Bayesian 泛化上界。Chérief 等[120]通过分析变分自编码器 (Variational Auto-Encoders, VAE) 重构误差的 PAC-Bayesian 泛化上界，研究了经典 VAE 目标的正则化效应。Mbacke 等[121]建立了更多 VAE 的 PAC-Bayesian 上界，Mbacke 等[122]则研究了对抗生成模型 (Generative Adversarial Models)。Haddouche 等[123]考虑了具有假设依赖取值范围的损失，并基于自界定函数建立了此类损失的上界。Haddouche 等[124]基于超鞅 (Supermartingales) 技术构建了重尾损失函数的上界。Haddouche 等[125]和 Viallard 等[126]提出了基于 Wasserstein 距离的 PAC-Bayesian 泛化上界，这些上界适用于无界损失，并可用作训练时的优化目标。

1.3.4 基于超样本技术的信息论泛化上界

以上相关工作中涉及的多数结果均需要满足绝对连续性假设，以确保基于 Donsker-Varadhan 变分表示的互信息度量的有界性。对于期望意义下的泛化上界，由 Bu 等[54]引入的随机子集或个体样本技术可在一定程度上缓解此问题，但其根源仍然存在：对于连续型数据分布，单个训练样本所携带的 Shannon 信息量可能是无界的。为此，Steinke 等[32]引入了基于超样本技术的泛化分析方法，其思想可追溯至 Audibert[39]和 Catoni[31]的工作，原本被用于减小 PAC-Bayesian 泛化上界的方差。在超样本框架中，泛化上界不再依赖训练数据与假设间的互信息，而是由假设和用于划分训练与测试数据的超样本变量间的互信息刻画，将学习算法的输入由连续型变量转换为离散型变量，以保证绝

对连续性始终得到满足。

Steinke 等[32]的工作贡献了该框架下的众多泛化分析结果，包括期望意义下基于条件互信息 (Conditional Mutual Information, CMI) 实现快速收敛率的泛化误差上界，以及面向无界损失函数的拓展上界。Hellström 等[51]推广了基于通用凸函数和二值 KL 散度的泛化上界。Haghifam 等[55]探索了分解和随机子集技术在超样本框架下的应用，这些结果在后续工作[50,57]中得到进一步完善。Hellström 等[87]与 Grünwald 等[127]继而探索了基于超样本技术的 PAC-Bayesian 泛化上界，该结果在 Bernstein 条件下给出了快速收敛率的、CMI 风格的 PAC-Bayesian 上界。Hellström 等[87]推广了超样本设定下的单次抽样上界，其可视为 Esposito 等[102]结果在 CMI 设定下的扩展。

Steinke 等[32]的工作指出，使用假设本身捕获的信息量作为泛化误差的衡量存在潜在性缺陷。例如，深度学习网络中通常存在许多对称性，因此不同的假设可以表达完全相同的预测函数，这种对称性却无法被 CMI 泛化上界捕捉。理想情况下，人们希望得到与神经网络预测值或损失值直接相关的泛化上界。基于此种考虑，Steinke 等[32]首次提出了评估条件互信息 (Evaluated CMI, e-CMI) 的概念，其直接使用网络输出的损失值构建信息论度量。事实证明，基于 e-CMI 或类似度量的泛化上界相比传统 CMI 框架显著增强了其紧致性。此种新方法保证了任何在超样本上给出相同损失值的假设均被视为等同，从而获得了捕捉网络对称性的能力。基于 e-CMI 的思想，后续工作[51,58,128-130]进一步拓展了基于网络预测值或损失值的泛化度量，推导了更紧的泛化误差上界。其中，Haghifam 等与 Rammal 等同时提出了基于留一法 (Leave-one-Out) 的 CMI 拓展泛化分析框架。Wang 等[131]将信息论泛化方法与算法稳定性技术相结合，改进了特定随机凸优化问题的泛化上界。Sachs 等[132]推导了基于算法依赖 Rademacher 复杂度的泛化上界，其在概念上与 CMI 框架相似。Sefidgaran 等[133]结合信息瓶颈和最小描述长度原则，给出了表示学习的泛化上界。

1.3.5 面向深度学习模型的信息论泛化上界

对于深度学习模型，一类被广泛使用的学习算法是基于随机梯度迭代的算法，其通过逐步更新假设以期收敛到具备良好泛化能力的最终结果。此类算法的一个经典实例是随机梯度下降算法 (Stochastic Gradient Descent, SGD)，其通过训练损失函数对网络参数的负梯度逐步更新当前假设，该负梯度由称为学习率的超参进行缩放。在现代机器学习场景中，深度神经网络占据了举足轻重的地位，其训练任务通常通过 SGD 算法及其变体完成。其中，一类对于信息论泛化分析尤为重要的变体称为随机梯度 Langevin 动力学 (Stochastic Gradient Langevin Dynamics, SGLD) 方法。SGLD 是一类引入了随机高斯噪声的 SGD 算法变体，使其尤其适用于信息论泛化分析。

Pensia 等[134]首次结合基于训练轨迹的互信息链式分解以及有限二阶矩高维随机变

量的 Shannon 熵上界，得到了面向 SGLD 算法的信息论期望泛化误差上界。基于其思想，众多后续工作[33,54-55,135-140]从各个方面对此上界进行了进一步改进。通过在 SGD 算法优化轨迹中引入辅助高斯噪声以与 SGLD 算法轨迹相关联，Neu 等[34]与 Wang 等[141]证明了 SGD 的信息论泛化上界。然而，Haghifam 等[142]指出，目前的信息论技术不足以获得随机凸优化问题的最优最小最大收敛速率。此外，Neyshabur 等[143]与 Bartlett 等[144]基于神经网络的权重范数推导泛化上界。Neyshabur 等[145]基于 PAC-Bayesian 视角，通过神经网络对参数扰动的鲁棒性，得到了可显式评估的基于相对熵的去随机化上界。Foret 等[146]与 Tsuzuku 等[147]探讨了 PAC-Bayesian 上界与平坦度的联系。Banerjee 等[148]探索了另一种引入噪声的 SGD 变体。Pitas[149]在神经网络 PAC-Bayesian 上界中引入高斯后验。Dziugaite 等[150]建立了 PAC-Bayesian 与 Entropy-SGD 的联系。

在无限网络宽度与特定初始化条件下，神经网络可通过高斯过程[151]描述 (Neural Network Gaussian Process, NNGP)。Pérez 等[152]将 PAC-Bayesian 上界与 NNGP 相结合，论证了神经网络所学习的函数在某种意义上是简单的，而这种简单性是其泛化能力的来源。Bernstein 等[153]通过类似方法推导了解析泛化上界。在特定损失函数和适当学习率选择下，训练期间无限宽度神经网络的演变可通过神经切线核 (Neural Tangent Kernel, NTK)[154]描述。Shwartz 等[155]基于 NTK 解析了神经网络相关信息度量。后续研究[156-158]将 NTK 拓展到通过优化 PAC-Bayesian 上界训练的网络，Wang 等[159]则探索了其与信息瓶颈的联系。

Viallard 等[160]基于 PAC-Bayesian 框架分析了一种特定两阶段神经网络的训练过程。Rivasplata 等[161]讨论了一类通过最小化 PAC-Bayesian 上界训练随机神经网络的方法。Letarte 等[162]研究了具有二值激活函数的神经网络，通过 PAC-Bayesian 上界制定其训练框架并建立泛化保证。Biggs 等[163]讨论了随机神经网络的集成方法 (Ensemble Method)，建立了可微分的 PAC-Bayes 目标。Biggs 等[164]通过数据依赖先验推导了浅层神经网络的去随机化 PAC-Bayesian 上界。Zantedeschi 等[165]通过 PAC-Bayesian 上界学习随机多数投票 (Stochastic Majority Votes)，而 Nagarajan 等[166]通过噪声抗性得到了去随机化的 PAC-Bayes 上界。Tinsi 等[167]基于高斯先验的 PAC-Bayesian 上界，得到了特定浅层聚合神经网络的可计算上界，而 Clerico 等[168]推导了一种无需代理损失的随机神经网络训练算法。Jin 等[169]基于权重扩展讨论了 Dropout 对于 PAC-Bayesian 泛化上界的影响。Liao 等[170]基于 PAC-Bayes 方法推导了图神经网络的泛化上界，Viallard 等[171]和 Xiao 等[172]则推导了对抗鲁棒性的上界。

1.4 研究内容

本论文聚焦于信息论视角下随机学习算法的泛化理论研究，针对目前信息论泛化理论的发展瓶颈，于现有工作基础之上继续探索发展基于信息论的泛化理论分析技术，

深入研究传统有监督学习、分布外领域泛化、无监督对比学习等经典学习范式的泛化理论，突破当前泛化分析结果在可计算性、紧致性以及适用范围等方面的局限性，构建可计算性强、估计精度高、适用范围广的信息论泛化分析理论框架，整体研究框架如图 1-1 所示。本论文的主要研究内容可划分为：

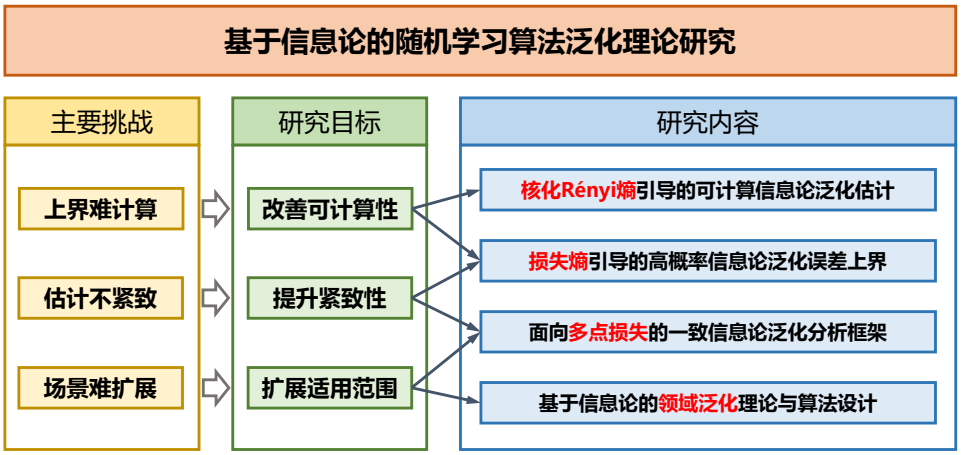


图 1-1 论文整体研究框架图

（1）核化 Rényi 熵引导的可计算信息论泛化估计

信息论相关泛化分析方法已在刻画随机迭代学习算法的泛化能力方面取得良好进展。然而，由于高维信息度量的不可计算性，此类上界在实际场景中往往难以量化计算，需要进一步建立相关信息量的上界估计以获得可计算的泛化误差上界。这一额外步骤引入了过度估计问题，导致其估计值严重偏离真实的误差值，仅仅能够粗略反映泛化误差的变化趋势，无法进行精准预测。本论文利用矩阵 Rényi 熵的核函数投影思想，将其拓展至无限样本情形，依此构建新型信息度量准则核化 Rényi 熵，继而推导面向 SGLD 和 SGD 等随机梯度迭代算法的可计算泛化误差上界。进一步地，针对矩阵 Rényi 熵的朴素算法计算效率低下的问题，本论文结合矩阵迹估计以及多项式近似技术，建立相关信息度量快速近似算法，并通过最优多项式近似理论，证明该算法复杂度达到理论最优收敛阶，为此类泛化上界的可计算性提供理论保障。

（2）损失熵引导的高概率信息论泛化误差上界

基于信息论的泛化分析是研究传统有监督学习框架下深度学习模型泛化能力的有效途径之一。然而，现有基于信息论的泛化理论多针对期望情形下的泛化误差刻画，缺乏高概率情形下的泛化理论探讨。同时，其推导过程依赖于高维随机变量的互信息度量，在实际应用中难以量化计算，无法准确地反映模型的实际泛化能力。本论文基于损失熵设计新型低维信息度量，并依此构建可计算的高概率泛化误差估计，分别探讨数据无关和数据依赖场景下的泛化误差度量准则，发掘损失熵度量与模型泛化误差间的强相关性，指出影响模型泛化能力的关键因素，并基于深度表示学习中的信息压缩思想优化现有泛化上界的紧致性，建立更为精确的深度学习泛化理论上界，为有监督深

度学习模型的设计、优化提供理论指导。

（3）面向多点损失的一致信息论泛化分析框架

近期，CLIP 等多模态对比学习大模型的兴起引发了国内外学者对于多点学习机制泛化性能的研究热潮。与传统的有监督单点学习机制不同，由于多点损失函数的引入，评估深度学习模型的性能不再通过独立同分布的单个样本，转而通过成对的样本子集进行。由此，目前面向单点损失函数构建的泛化分析理论不再适用于对比学习、度量学习、排序算法等多点学习场景。同时，基于算法的一致收敛性或稳定性理论推导的泛化误差上界通常依赖于假设空间的复杂性或损失的 Lipschitz 连续性、光滑性、凸性等强假设，从而难以应用于现代深度神经网络的泛化分析。本论文从基于信息论的单点学习泛化理论出发，克服由引入多点损失函数导致的样本独立同分布性质丢失与维度爆炸等挑战，通过发展新型的信息论样本解耦技术，构建面向多点损失的一致泛化分析框架，为对比学习及考虑高阶样本耦合等多点学习机制提供理论指导。

（4）基于信息论的领域泛化理论与算法设计

领域泛化是分布外泛化研究领域的代表性子任务之一，致力于解决实际采样数据中存在的分布偏移问题。具体而言，机器学习模型的训练数据通常由多个数据源构成，这些数据根据其来源不同存在分布上的差异，将导致模型过拟合到某些数据分布的独有特征，从而在新数据上泛化效果下降。领域泛化通过将训练集划分为不同领域，每个领域对应一个独有的数据分布，旨在学习不同分布间的共有信息，从而在新领域上保持其泛化性能。然而，现有理论多将领域泛化视为平均或最坏情形下的优化问题，此类理论或对分布外数据不具备鲁棒性，或将导致解空间过分受限。本论文基于信息论方法构建领域泛化的概率学习理论，探讨影响模型在源域与目标域上实现泛化的关键信息度量，基于理论分析解释目前领域泛化算法成功的内在机理，并据此设计新型高效、稳健的领域泛化学习算法，提升深度学习模型的分布外泛化性能。

1.5 主要贡献与创新之处

本论文属于统计机器学习基础理论与方法领域的重要研究方向，是人工智能与信息论、统计学交叉融合的前沿研究课题。相较于国内外现有研究工作，本项目的主要贡献与创新之处可归纳如下：

（1）可计算信息论泛化度量与近似算法。本论文突破了现有基于传统 Shannon 熵信息度量的泛化分析体系，结合新型可计算信息度量矩阵 Rényi 熵，研发了可计算的信息论泛化度量：核化 Rényi 熵，克服了传统 Shannon 熵在高维情形下难以计算的根本性缺陷。本论文基于核化 Rényi 熵构建了 SGD/SGLD 等随机梯度迭代算法的泛化上界，建立了泛化误差与轨迹梯度协方差间的关联。进一步地，本论文面向矩阵 Rényi 熵研发了

理论最优快速近似算法，为相关信息论泛化上界的可计算性提供切实保障。

（2）低维信息论泛化度量与高概率误差估计。一方面，本论文克服了目前基于假设互信息或超样本方法的信息论泛化上界普遍存在的不可计算、紧致性不足等缺陷，提出新型低维信息论泛化度量：损失熵。该度量仅包含一维随机变量，可通过核密度估计、分箱方法等直接近似计算，为真实神经网络的泛化能力提供数值估计结果。另一方面，本论文结合基于典型子集及超样本集的泛化分析技术，构建了基于损失熵的高概率泛化误差上界估计，可精确刻画随机算法的实际泛化性能，为复杂随机学习算法的泛化分析提供了新思路和新方法。

（3）多点学习泛化理论的信息论分析框架。本论文突破了基于一致收敛以及算法稳定性的双点、三点学习理论分析定式，从信息论角度建立全新的多点学习泛化理论分析框架。通过发展基于互信息分解的多点泛化误差解耦技术以及基于超样本方法的互信息降维分析技术，构建面向普适随机学习算法的多点学习泛化上界以及基于低维互信息的可计算泛化上界，进一步针对 SGD/SGLD 等随机梯度迭代优化算法构建基于轨迹梯度协方差的多点学习泛化上界，据此分析影响泛化的关键因素，为设计新型高效的多点学习算法提供理论基础。

（4）领域泛化学习理论的信息论基础。本论文突破了现有领域泛化理论框架仅关注于平均或最坏泛化情形的局限性，构建了信息论视角下的概率领域泛化分析框架。与平均情形下的理论相比，该框架具备更强的泛化约束性，能够提供切实的泛化性能保障；与最坏情形的理论分析相比，该框架更贴近实际学习过程并具有更强的可解释性，通过引入领域分布先验知识避免过度保守的泛化误差估计。同时，通过理论分析发现域间梯度对齐与域间特征对齐共同构成了解决领域泛化问题的充分条件，据此设计基于域间分布对齐的领域泛化算法，进一步提升领域泛化模型的分布外泛化性能。

1.6 论文组织结构

本论文面向基于信息论的泛化理论分析，针对现有工作在可计算性、紧致性以及适用范围方面的局限性展开研究，各章内容组织的具体安排如下：

第 1 章为绪论。本章首先介绍了统计学习理论中泛化分析问题的研究背景，简述了假设空间复杂度、算法稳定性等不同泛化分析方法的发展历程，分析了目前研究所面临的挑战。进一步地，以信息论视角下的泛化分析理论与工具为主线，详细介绍了国内外相关研究的发展现状。最后，介绍了本论文的研究内容、主要贡献与创新之处。

第 2 章为核化 Rényi 熵引导的可计算信息论泛化估计。本章针对现有基于传统 Shannon 熵信息度量的泛化误差上界难以量化计算的问题，研发了新型可计算信息论泛化度量：核化 Rényi 熵，并基于此度量构建了多种随机迭代算法的泛化上界，探索了泛

化与轨迹梯度协方差之间的联系；同时，基于矩阵迹估计与多项式近似方法设计了该度量的理论最优快速近似算法，为相关泛化上界的可计算性提供理论保障。

第 3 章为损失熵引导的高概率信息论泛化误差上界。本章针对现有基于信息论的高概率泛化上界难以计算、估计不紧致的问题，提出了新型低维信息论泛化度量：损失熵。该度量仅包含一维随机变量，从根本上解决了相关泛化误差上界的可计算性难题；同时，结合基于典型子集及超样本集的泛化分析技术，分别从数据无关与数据依赖两种问题设定入手，提升了多种现有高概率泛化误差上界的紧致性。

第 4 章为面向多点损失的一致信息论泛化分析框架。本章针对现有面向单点损失构建的泛化分析理论难以拓展至多点学习场景的问题，通过基于互信息分解的多点泛化误差解耦技术以及基于超样本方法的互信息降维分析技术，突破了多点学习泛化分析中的非独立同分布损失以及维度爆炸难题，构建了面向普适随机学习算法的多点学习泛化上界，从全新的信息论视角建立了统一的多点学习泛化分析框架。

第 5 章为基于信息论的领域泛化理论与算法设计。本章针对现有基于数据独立同分布假设的有监督泛化分析框架难以拓展至分布外泛化场景的问题，突破了目前领域泛化理论在优化目标刻画上的局限性，构建了信息论视角下的概率领域泛化分析框架，通过理论分析发现解决领域泛化问题的一种充分条件，并据此设计新型领域泛化算法以进一步提升其分布外泛化性能，填补了当下信息论泛化理论体系的一大空缺。

第 6 章为结论与展望。本章总结了本论文对于现阶段挑战的解决方案，阐述了可进一步改进或拓展的未来方向，并对下一阶段的研究工作进行规划与展望。

1.7 基本符号与概念

本论文中，随机变量使用大写字母（X）表示，其具体取值使用小写字母（x）表示，取值空间使用花体字母（$\mathcal{X}$）表示。随机变量 X 的概率分布使用 P_X 表示，给定 Y 时 X 的条件概率分布使用 $P_{X|Y}$ 表示，在特定取值下的条件概率分布则使用 $P_{X|Y=y}$（或 $P_{X|y}$）表示。类似地，$\mathbb{E}_X$、Var_X 和 Cov_X 分别表示对随机变量 $X \sim P_X$ 取期望、方差与协方差矩阵。设 $H(X)$ 为随机变量 X 的 Shannon 熵（对于连续型随机变量则为微分熵），$D(P \,\|\, Q)$ 为概率分布 P 相对于 Q 的 KL 散度。另外定义 $d(p \,\|\, q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ 为二值 KL 散度。设 $I(X;Y)$ 为随机变量 X 与 Y 的互信息，$I(X;Y|Z)$ 为给定 Z 时随机变量 X 与 Y 的条件互信息。进一步地，$F(X;Y) = D(P_{X,Y|z} \,\|\, P_{X|z}P_{Y|z})$ 表示解构互信息。本论文使用 $\mathbb{W}(\cdot,\cdot)$ 表示 Wasserstein 距离，log 为自然对数函数。

设 $\mathcal{Z}$ 为样本所在的实例空间，μ 为 $\mathcal{Z}$ 上的未知数据分布。对于传统有监督学习场景，实例空间 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 可进一步拆分为输入空间 $\mathcal{X}$ 与标签空间 $\mathcal{Y}$。设 $\mathcal{W} \subseteq \mathbb{R}^d$ 为假设空间。训练数据集 $\mathbf{Z} = \{Z_i\}_{i=1}^n \in \mathcal{Z}^n$ 通过从 μ 中独立同分布采样构建。学习算法 $\mathcal{A}$

以 $\mathbf{Z}$ 为输入，依据特定条件分布 $P_{W|\mathbf{Z}}$ 输出假设 $W \in \mathcal{W}$。设 $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$ 为损失函数。对于给定假设 $w \in W$，其总体风险定义为：

$$L(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]. \tag{1-2}$$

学习算法的最终目标是寻找能够最小化总体风险的假设 w。但由于数据分布 μ 未知，在实践中通常转而求解训练数据集 $\mathbf{Z}$ 上的经验风险，定义为：

$$L_{\mathbf{Z}}(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i). \tag{1-3}$$

给定假设 $w \in W$，其在训练数据集 $\mathbf{Z}$ 上的泛化误差定义为：

$$\overline{\mathrm{gen}}(w, \mathbf{Z}) \triangleq L(w) - L_{\mathbf{Z}}(w). \tag{1-4}$$

此外，定义 $L = \mathbb{E}_W[L(W)]$ 为期望总体风险，$L_n = \mathbb{E}_{W,\mathbf{Z}}[L_{\mathbf{Z}}(W)]$ 为期望经验风险。期望意义下学习算法 $\mathcal{A}$ 的泛化误差定义为：

$$\overline{\mathrm{gen}} \triangleq L - L_n = \mathbb{E}_{W,\mathbf{Z}}[L(W) - L_{\mathbf{Z}}(W)]. \tag{1-5}$$

其中期望取自于 $(W, \mathbf{Z})$ 的联合分布（即 $P_{W|\mathbf{Z}} \otimes \mu^n$）。

2 核化 Rényi 熵引导的可计算信息论泛化估计

2.1 引言

过参数化现象是现代深度神经网络模型训练过程中的一种常见现象，即其能够在拟合全部训练数据的同时，依旧表现出良好的泛化能力。传统统计学习理论将机器学习模型的泛化能力归因于假设空间复杂度，而 VC 维度、Rademacher 复杂度等经典复杂度度量往往对于模型规模较为敏感[173]，无法对大规模深度神经网络的过参数化现象做出有效解释。近期研究发现，学习算法的选择对于神经网络的泛化能力具有显著影响[144,174]，这引发了国内外学者对于不同学习算法理论属性的研究热潮[34,134,138,175]。

随机梯度下降算法已成为训练现代深度学习模型的事实标准。其思想与实现过程虽然十分简单，但却能够在复杂非凸优化问题中表现出良好的泛化性能[176]。这激发了后续面向深度学习优化算法的系列泛化理论研究工作。第一类方法采用一致稳定性概念，始于 Hardt 等[174]对期望意义下算法收敛性的探索，类似思想随后在大量后续研究中得到进一步拓展[177-180]。另一类方法将深度神经网络的泛化能力与特定关键信息论度量相关联[28]，在分析随机迭代学习算法方面同样展示出巨大潜力。Pensia 等[134]首次研究了 SGLD 算法的信息论泛化理论上界，其结果在后续研究中得到改进[33,138]；Neu 等[34]通过引入辅助优化轨迹并在其中加入虚拟高斯噪声，首次为 SGD 算法建立了信息论泛化上界，其结果在 Wang 等[141]的工作中进一步收紧。在稳定性与信息论视角之外，还出现了 PAC-Bayesian[116,145]以及模型压缩[181-182]视角下的泛化分析研究工作。

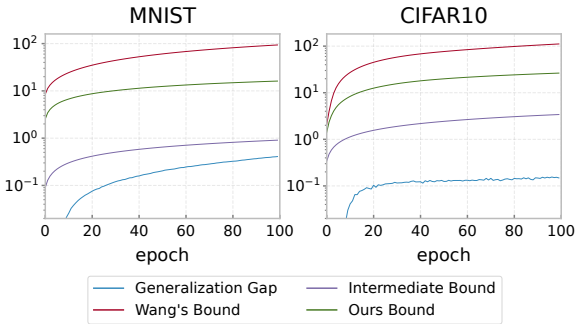


图 2-1 不同信息论泛化误差上界的数值对比

虽然目前针对深度学习算法泛化能力的理解与解释工作已取得了阶段性成果，这些泛化理论上界在前提假设或紧致性方面仍存在严重的局限性，难以应用于大规模深度神经网络的泛化性能分析。首先，基于一致稳定性的泛化理论通常需要其参数化损失函数满足 Lipschitz 连续性、平滑性等假设条件[174,178]或（强）凸性、Polyak-Lojasiewicz（PL）条件等全局最优解的存在性假设[175,183]以保证其收敛性，这些假设条件在实际网

络中往往难以满足。其次，基于信息论的泛化理论结果虽然不依赖于损失函数的相关强假设，但却通常依赖于假设空间的维度，往往导向严重过度估计的泛化理论上界。如图 2-1 所示，真实泛化误差值与现有信息论泛化上界间存在 10^2 到 10^3 倍的差异[138]。此外，高维随机变量相关信息度量的不可计算性为这些上限的进一步改进带来了额外阻碍。无论从数值角度或是理论角度，均无法对泛化上界推导过程中某一中间步骤的紧致性进行有效分析，故而难以得知目前泛化上界推导的瓶颈所在。

针对 Shannon 熵及其推广 Rényi 熵[184]对于高维随机变量难以量化计算的问题，Giraldo 等引入了基于矩阵的 Rényi 熵[185]，作为一类可直接通过给定数据样本量化计算的替代度量。本章探索了矩阵 Rényi 熵在无限样本下的拓展定义，称为核化 Rényi 熵，并依此构建了适用于随机迭代学习算法的信息论泛化上界。核化 Rényi 熵继承了原始 Shannon 熵的多项优良性质，同时对于任意维度的随机变量，均可直接通过给定采样数据量化计算。随后，本章基于核化 Rényi 熵分析了期望意义下随机迭代算法的泛化误差上界，其中的关键信息度量均可通过对训练过程抽样进行量化与可视化。基于此类可视化结果，分析了目前信息论泛化分析方法的瓶颈所在，并在现有结果基础之上进行了针对性优化。例如，Wang 等[138]在先前工作中通过梯度轨迹方差构建了关键信息度量的上界估计，如图 2-1 所示，这类结果（红色曲线）严重高估了相关信息量的实际值（紫色曲线），存在 10 到 10^2 倍以上的差异。这些观察结果启发了进一步考虑梯度向量不同维度间的相关性，基于梯度轨迹协方差构建泛化误差上界以改进其紧致性（绿色曲线）。这种改进方法同样适用于 Pensia 等[134,141]的工作。同时，针对矩阵 Rényi 熵计算复杂度 $O(n^3)$（n 为样本数量）较高的问题，从随机数值线性代数角度出发，结合矩阵迹估计与多项式近似技术设计了相关快速近似算法，将总体计算复杂度降低至 $O(n^2sm)$（$s, m \ll n$），并证明了相关参数 s 和 m 已经达到最优收敛阶。

总体而言，本章的主要贡献包括：（1）提出了基于 Hilbert 空间算子表示的核化 Rényi 熵。与经典 Shannon 熵不同，核化 Rényi 熵对于任意维度的随机变量均可直接计算，同时依然兼容现有的信息论泛化分析框架。（2）基于核化 Rényi 熵为 SGD/SGLD 等随机迭代学习算法建立了可计算的信息论泛化上界，并基于相关可视化结果指出了目前泛化理论的优化瓶颈，通过引入梯度轨迹协方差提升了现有泛化上界的紧致性。（3）基于矩阵迹估计与多项式近似技术开发了针对矩阵 Rényi 熵的快速近似算法，并证明其已达到理论最优复杂度，为相关泛化上界的可计算性提供切实保障。

2.2 Rényi 熵及拓展信息度量

给定连续型随机变量 X 及其概率密度函数 $p(x)$，其对应的 Shannon 熵度量定义为：

$$H(X) \triangleq -\int_{\mathcal{X}} p(x)\log p(x)\,\mathrm{d}x. \tag{2-1}$$

Rényi 熵是一类拓展熵度量方法，其通过超参数 α（$\alpha > 0$ 且 $\alpha \neq 1$）包含了一系列不同的熵度量，包括 Shannon 熵（$\alpha \to 1$）、最小熵（$\alpha \to \infty$）和碰撞熵（$\alpha = 2$）等：

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} p^\alpha(x)\, \mathrm{d}x, \tag{2-2}$$

通过以上定义可知，Shannon 熵与 Rényi 熵均依赖于随机变量的概率分布，而高维随机变量的概率密度估计十分困难，限制了其在神经网络等高维场景下的应用。为此，Giraldo 等提出了一种替代熵度量方法，其能够直接通过给定数据样本量化计算：

定义 2.1 ([185]，命题 3.1) 给定实值正定无限可分核函数 $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ 和数据样本 $\{x_i\}_{i=1}^n \subset \mathcal{X}^n$，并计算 Gram 矩阵 $K_{ij} = \kappa(x_i, x_j)$，则 α 阶矩阵 Rényi 熵可定义为：

$$S_\alpha(A) = \frac{1}{1-\alpha} \log \mathrm{tr}(A^\alpha) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} \lambda_i^\alpha(A), \tag{2-3}$$

其中 $A_{ij} = \frac{1}{n} \frac{K_{ij}}{\sqrt{K_{ii} K_{jj}}}$ 为归一化核矩阵，$\lambda_i(A)$ 表示矩阵 A 的第 i 个特征值。

矩阵 Rényi 熵也可进一步拓展至 L 个随机变量的联合熵：

$$S_\alpha(A_1, \cdots, A_L) = S_\alpha\left(\frac{A_1 \circ \cdots \circ A_L}{\mathrm{tr}(A_1 \circ \cdots \circ A_L)} \right), \tag{2-4}$$

其中 $A_1, \cdots, A_L$ 为每个随机变量对应的归一化的核矩阵，$\circ$ 为 Hadamard 乘积。进一步地，可定义基于矩阵 Rényi 熵的条件熵、互信息等信息度量：

$$S_\alpha(A_1, \cdots, A_k | B) = S_\alpha(A_1, \cdots, A_k, B) - S_\alpha(B), \tag{2-5}$$

$$I_\alpha(\{A_1, \cdots, A_k\}; B) = S_\alpha(A_1, \cdots, A_k) - S_\alpha(A_1, \cdots, A_k | B), \tag{2-6}$$

可见，上述基于矩阵的熵度量方法避免了高维随机变量的概率密度估计，从而能够在高维场景中保持其可计算性。

2.3 新型熵度量法则：核化 Rényi 熵

本节通过将矩阵 Rényi 熵从有限样本扩展到无限样本情形，引入了核化 Rényi 熵，从而能够从其概率密度函数出发，直接分析相关变量的信息熵而不受实际采样过程影响。这种定义继承了原始 Shannon 熵的优良性质，同时仍然可以通过简单的数据采样直接量化计算。其仍遵循 Giraldo 等[185]的再生核 Hilbert 空间表示框架，同时对所用的核函数做特定假设：

假设 2.2 给定再生核函数 $\kappa(x, x') = \langle \varphi(x), \varphi(x') \rangle$，其中 φ: $\mathcal{X} \mapsto \mathcal{H}$ 为对应的特征映射关系。假设 κ 满足以下条件：

（1） 归一化：对于任意 $x \in \mathcal{X}$，有 $\kappa(x,x) = 1$；

（2） 平移不变性：存在 $f\colon \mathbb{R}^+ \mapsto \mathbb{R}^+$，使得 $\kappa(x,x') = f(\|x - x'\|)$；

（3） L_2 可积：对于任意 $x \in \mathcal{X}$，有 $\int_{\mathcal{X}} \kappa^2(x,x')\,\mathrm{d}x' < \infty$.

给定随机变量 $X \in \mathcal{X}$，定义线性算子 $G_X : \mathcal{H} \mapsto \mathcal{H}$，$G_X(f) \triangleq \mathbb{E}_X[\varphi(x)\langle\varphi(x), f\rangle]$。当核函数 κ 满足归一化条件时，容易验证 $\mathrm{tr}(G_X) = 1$。从而，G_X 的特征值集合构成一个随机变量的概率分布，可作为变量 X 概率分布的自然密度估计。

定理 2.3 给定连续型随机变量 $X \in \mathcal{X}$ 及其概率密度函数 p_X，则上述线性算子 G_X 满足

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha} \log \mathrm{tr}(G_X^{\alpha}) = -\mathrm{tr}(G_X \log G_X) = -\iint_{\mathcal{X}^2} p_X(x) \log p_X(x') \kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x'. \tag{2-7}$$

上述定理直接隐含了如下的 $\alpha \to 1$ 阶核化 Rényi 熵定义：

定义 2.4 给定连续型随机变量 X 及其概率密度函数 p_X，则随机变量 X 的 $\alpha \to 1$ 阶核化 Rényi 熵可定义为

$$S_1(X) \triangleq -C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log p_X(x') \kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x'. \tag{2-8}$$

其中 $C_\kappa = 1/\int_{\mathcal{X}} \kappa^2(0,x)\,\mathrm{d}x > 0$ 是归一化因子，以确保该核函数平方积分为 1。

与难以有效应对高维随机变量的经典 Shannon 熵相比，上述核化 Rényi 熵不受变量维度影响，可直接量化计算。为此，可从概率分布 P_X 中随机采样 m 个数据点 $\{x_i\}_{i=1}^m$，并定义 $\widehat{G}_X : \mathcal{H} \mapsto \mathcal{H}$ 为 G_X 的经验版本：

$$\widehat{G}_X(f) \triangleq \frac{1}{m} \sum_{i=1}^{m} \varphi(x_i)\langle\varphi(x_i), f\rangle. \tag{2-9}$$

可以验证 $\widehat{G}_X$ 是 G_X 的无偏估计，依此可得核化 Rényi 熵的有限样本近似算法：

定理 2.5 给定独立同分布样本数据 $\{x_i\}_{i=1}^m$，计算核矩阵 $K \in \mathbb{R}^{m \times m}$：$K_{ij} = \frac{1}{m}\kappa(x_i, x_j)$。则在置信度 $1-\delta$ 下，有

$$|S_1(X) - \widehat{S}_1(X)| \leq \frac{9C_\kappa \sqrt{2 \log \frac{2}{\delta}}}{\sqrt[3]{m}}, \tag{2-10}$$

其中 $\widehat{S}_1(X) = -C_\kappa \mathrm{tr}(K \log K)$。

注意到，上述近似误差上界仅涉及采样样本数量 m 而与随机变量 X 的维度无关。因此，核化 Rényi 熵对于高维随机变量仍可通过有限样本直接计算，这一特性在分析现代深度神经网络的泛化能力时具有显著优势。同时，定义 2.4 可通过取 $\kappa = \kappa_X \otimes \kappa_Y$ 作为联合分布 $P_{X,Y}$ 的核函数，从而拓展至多个随机变量的联合熵。基于这种定义，可相应地导出核化 Rényi 散度与互信息的数学定义：

定义 2.6 给定 $\mathcal{X}$ 上的概率分布 P、Q 及其概率密度函数 p、q，则 P 相对于 Q 的核化

Rényi 散度定义为：

$$D_1(P \parallel Q) \triangleq C_\kappa \iint_{\mathcal{X}^2} p(x) \log \frac{p(x')}{q(x')} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x'. \tag{2-11}$$

定义 2.7 给定归一化核函数 κ_X、κ_Y，连续型随机变量 X、Y 及其概率密度函数 p_X、p_Y，则 X 与 Y 的核化 Rényi 互信息定义为：

$$I_1(X;Y) \triangleq C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2 \times \mathcal{X}^2} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x',y')}{p_X(x')p_Y(y')} \kappa_X^2(x,x') \kappa_Y^2(y,y') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \,\mathrm{d}y'. \tag{2-12}$$

由上述定义可见，Shannon 熵与核化 Rényi 熵的主要区别在于引入了核函数 κ。具体而言，当 κ 取 Dirac-Delta 函数时，上述定义可退化为原始 Shannon 熵。为了形式化表征其差异性，引入离差函数

$$u_X^\kappa(x) \triangleq C_\kappa \int_{\mathcal{X}} \left[\log p_X(x) - \log p_X(x') \right] \kappa^2(x,x') \,\mathrm{d}x', \tag{2-13}$$

及其期望形式

$$E_X^\kappa(p) \triangleq \left| \int_{\mathcal{X}} p(x) u_X(x) \,\mathrm{d}x \right|. \tag{2-14}$$

简便起见，将**期望离差** $E_X^\kappa(p_X)$ 简写为 $E_X^{\kappa\prime}$，并将 $E_X^\kappa(\hat{p}_X)$ 简写为 E_X^κ，其中

$$\hat{p}_X(x) \triangleq C_\kappa \int_{\mathcal{X}} p_X(x') \kappa^2(x,x') \,\mathrm{d}x'. \tag{2-15}$$

定理 2.8 设 $\kappa(x,x') = 1_{\|x-x'\|<c}$。假设概率密度函数 $p_X(\cdot)$ 满足：

（1） 连续性：给定任意 $x \in \mathcal{X}$，有 $\lim_{x' \to x} p_X(x') = p_X(x)$;

（2） 正值性：给定任意 $x \in \mathcal{X}$，有 $\lim_{x' \to x} p_X(x') > 0$;

则有 $\lim_{c \to 0} E_X^\kappa \to 0$，以及 $\lim_{c \to 0} E_X^{\kappa\prime} \to 0$.

上述定理表明，当选择的核函数拥有较为尖锐的峰值（即 c 较小）时，期望离差项 E_X^κ 与 $E_X^{\kappa\prime}$ 均趋于 0。正如定理 2.9 所示，设置 $c \to 0$ 对应于核化 Rényi 熵退化为原始 Shannon 熵的情形。上述连续性假设在 X 为连续型随机变量时容易满足。正值性假设也在 X 的取值范围截断于特定区间 $[a,b]$ 时自然满足，即若 $x \in [a,b]$ 则有 $p_X(x) > 0$，反之则有 $p_X(x) = 0$（例如图像数据总是截断于 $[0,255]$）。另一种情况是当 X 拥有长尾分布（例如高斯分布用于模型参数初始化）时，上述正值性假设同样得以满足。上述均为现代深度神经网络训练中的常见情形。

定理 2.9 设 $X, X' \in \mathcal{X}$、$Y \in \mathcal{Y}$ 与 $Z \in \mathcal{Z}$ 为连续型随机变量，其概率分布分别为 P_X、$P_{X'}$、P_Y 与 P_Z，则有

（1） $H(X) \le S_1(X) \le H(X) + E_X^{\kappa\prime}$.

（2） $D_1(P_X \parallel P_{X'}) \geq -E_X^{\kappa}$.

（3） $I_1(X;Y) = D_1(P_{X,Y} \parallel P_X \otimes P_Y) \geq 0$.

（4） $I_1(X;Y) = S_1(X) + S_1(Y) - S_1(X,Y)$.

（5） $I_1(X;Y|Z) = I_1(X;Y,Z) - I_1(X;Z)$.

（6） 设 X, Y, Z 形成 Markov 链 $X \to Y \to Z$，则 $I_1(X;Y) \geq I_1(X;Z)$ 且 $I_1(Y;Z) \geq I_1(X;Z)$。

上述定理表明核化 Rényi 熵继承了原始 Shannon 熵的基本性质，从而保证了其与现有信息论泛化分析框架的兼容性。性质 1 验证了当定理 2.8 中的 $c \to 0$ 时，核化 Rényi 熵退化为原始的 Shannon 熵。结合其他性质，这一结论也可拓展至核化 Rényi 散度和互信息。性质 2 表明，虽然核化 Rényi 散度不恒为正，但通过选择合适的核函数 κ，可保证其不显著小于 0。性质 4 与 5 表明，可通过近似计算多项核化 Rényi 熵以获得核化 Rényi 互信息的近似结果。性质 6 是数据处理不等式的核化 Rényi 熵变体。

以下分析将基于高斯核函数推导期望意义下的泛化误差上界，即

$$\kappa(x,x') = \exp\left(-\|x-x'\|_2^2/2\sigma_\kappa^2\right), \tag{2-16}$$

其中 σ_κ 为核宽度。值得注意的是，σ_κ 的取值需具体斟酌权衡：较小的 σ_κ 将使 $E_X^{\kappa} \approx 0$，对应于退化为 Shannon 熵的情形。然而如定理 2.5 所示，这同时将使归一化因子 C_κ 增大并导致较大的近似误差。如 Yu 等[186]所建议，实践中可根据所有成对数据点之间的前 10% 至 20% 欧氏距离选择 σ_κ 的取值。

2.4 基于核化 Rényi 熵的泛化误差上界

本节基于上述核化 Rényi 熵的定义与基本性质，结合当前信息论泛化分析框架推导了面向随机迭代学习算法的信息论泛化误差上界。Xu 等[28]的工作证明了在损失函数满足 σ-次高斯条件时，有如下的互信息泛化上界：

$$|\overline{\mathrm{gen}}| \leq \sqrt{\frac{2\sigma^2 I(W;\mathbf{Z})}{n}}, \tag{2-17}$$

由于 W 与 $\mathbf{Z}$ 均为高维随机变量，上述泛化上界在实际应用中无法量化计算。基于其基本思想，可证明上述泛化误差互信息上界同样适用于核化 Rényi 熵：

定理 2.10 假设 $\ell(w,Z)$ 对于任意 $w \in \mathcal{W}$ 均满足对于 $Z \sim \mu$ 的 σ-次高斯性，则

$$|\overline{\mathrm{gen}}| \leq \sqrt{\frac{2\sigma^2 \widehat{I}_1(W;\mathbf{Z})}{n}}, \tag{2-18}$$

$$\mathbb{E}_{W,\mathbf{Z}}[\overline{\mathrm{gen}}^2(W;\mathbf{Z})] \leq \frac{4\sigma^2(\widehat{I}_1(W;\mathbf{Z}) + \log 3)}{n}, \tag{2-19}$$

其中 $\widehat{I}_1(W;\mathbf{Z}) = I_1(W;\mathbf{Z}) + E_{W,\mathbf{Z}}^{\kappa}$。

上述定理提供了核化 Rényi 熵视角下的信息论泛化误差上界。如定理 2.8 所示，$E^{\kappa}_{W,\mathbf{Z}}$ 是 $(W,\mathbf{Z})$ 的联合分布相对于核函数 κ 的期望离差，且当 $\sigma_\kappa \to 0$ 时，该离差将趋于 0。值得注意的是，定理 2.10 通过输入—输出互信息 $I_1(W;\mathbf{Z})$ 同时建立了 $\overline{\mathrm{gen}}(W;\mathbf{Z})$ 期望与方差的理论上界，故可进一步结合集中不等式（如 Markov 或 Chebyshev 不等式）给出高概率的泛化误差上界估计结果。

以下将上述泛化分析结果应用于基于随机批次迭代学习算法的泛化性能分析。假设算法 $\mathcal{A}$ 共执行 T 步更新，其中 $W_0 \in \mathcal{W}$ 为初始参数向量。在第 t 步迭代中，学习算法将从训练数据集中随机选择一批数据点 $B_t \subset \mathbf{Z}$ 用于计算梯度下降方向：

$$g(w,B_t) \triangleq \frac{1}{|B_t|}\sum_{z\in B_t}\nabla_w \ell(w,z). \tag{2-20}$$

其更新规则可形式化为

$$W_t = W_{t-1} - \eta_t g(W_{t-1},B_t) + \xi_t, \tag{2-21}$$

其中 W_t 为第 t 步迭代时的参数向量，η_t 为学习率，$\xi_t \in \mathcal{W}$ 为与 W_{t-1} 和 B_t 均独立的随机噪声向量。可观察到 $W_0 \to W_1 \to \cdots \to W_T$ 构成了 Markov 链。

2.4.1 随机梯度 Langevin 动力学算法

SGLD 算法是经典 SGD 算法的变体之一，其通过在梯度迭代过程中加入随机噪声增强模型的泛化能力。一类常用的噪声分布是高斯噪声，即 $\xi_t \sim N(0,\sigma_t^2 I_d)$。Pensia 等[134]发现在所有固定方差的随机变量中，高斯噪声具有最大的 Shannon 熵，因而可导向最紧的上界估计。Pensia 等的泛化估计结果如下所示：

引理 2.11 ([134]，定理 1) 设 W_T 为 SGLD 算法在 T 步更新后得到的参数向量，则

$$I(W_T;\mathbf{Z}) \le \sum_{t=1}^{T}\frac{d}{2}\log\left(\frac{\eta_t^2 L}{d\sigma_t^2}+1\right), \tag{2-22}$$

其中 $L = \max_{w\in\mathcal{W},z\in\mathcal{Z}}\|g(w,z)\|_2^2$.

若损失函数 $\ell(w,z)$ 对于 w 满足 Lipschitz 连续性，则引理 2.11 中的常数 L 可替换为损失函数的 Lipschitz 常数。Wang 等[138]对其做出了进一步改进，移除了对于损失函数的 Lipschitz 假设：

引理 2.12 ([138]，定理 1) 在引理 2.11 的相同条件下：

$$I(W_T;\mathbf{Z}) \le \sum_{t=1}^{T}\frac{\eta_t^2 V_t}{2\sigma_t^2}, \tag{2-23}$$

其中 V_t 是第 t 步迭代时的**梯度方差**，定义为

$$V_t \triangleq \mathbb{E}_{W_{t-1},B_t}\Big[\|g(W_{t-1},B_t)-\mathbb{E}_{B_t}[g(W_{t-1},B_t)]\|_2^2\Big]. \tag{2-24}$$

上述泛化上界并未显式依赖于模型参数维度 d。然而，梯度方差 V_t 度量了随机梯度向量各个维度方差的总和，故而隐式依赖于 d。这种结果源自上界推导过程中使用了各向同性的高斯分布作为固定方差随机变量的 Shannon 熵上界。为此，考虑引入梯度向量不同维度之间的相关性，从而为 SGLD 算法提供严格更紧的泛化上界。

定理 2.13 在引理 2.11 的相同条件下：

$$I_1(W_T;\mathbf{Z}) \le \sum_{t=1}^{T} I_1(W_t;B_t|W_{t-1}) \le \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t+I\right| + E^{\kappa}_{W_t|W_{t-1}}\right), \tag{2-25}$$

$$I(W_T;\mathbf{Z}) \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t+I\right|, \tag{2-26}$$

其中 $\mathcal{V}_t=\mathrm{Cov}[g(W_{t-1},B_t)]$ 为**梯度协方差**矩阵，$|\cdot|$ 表示矩阵的行列式。

上述定理指出，通过核化 Rényi 互信息 $I_1(W_T;\mathbf{Z})$ 可建立由梯度协方差矩阵行列式作为主要度量的泛化上界。当取极限 $\sigma_\kappa\to 0$ 时，可导出式 (2-26) 中的 Shannon 互信息上界。上述式 (2-25) 中的核化 Rényi 信息量均可通过优化轨迹采样直接计算，进一步结合定理 2.10 即可得到期望意义下 SGLD 算法的泛化误差上界。

定理 2.14 给定如上定义的 $\mathcal{V}_t$、V_t 与 L，令 $\{c_i\}_{i=1}^r$ 为 $\{n\}$ 的一种无重复划分，即 $c_1\cup\cdots\cup c_r=\{n\}$，且对于任意 $i\neq j$ 有 $c_i\cap c_j=\Phi$。设 $\mathcal{V}_t^i$ 为行列均由 c_i 索引的 $\mathcal{V}_t$ 子矩阵，并定义

$$\theta_c(\mathcal{V})=\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}+I\right|,\quad \theta_v(V)=\frac{d}{2}\log\left(\frac{\eta_t^2 V}{d\sigma_t^2}+1\right), \tag{2-27}$$

$$\text{则有}\qquad \theta_c(\mathcal{V}_t)\le\sum_{i=1}^{r}\theta_c(\mathcal{V}_t^i)\le\theta_v(V_t)\le\theta_v(L). \tag{2-28}$$

上述定理中的 θ_c 与 θ_v 度量分别对应于定理 2.13 和引理 2.12 中的泛化上界。当神经网络规模 d 较大时，由于计算机内存受限，无法完整地计算整个协方差矩阵 $\mathcal{V}_t$。上述定理提出了 $\theta_c(\mathcal{V}_t)$ 的一种替代上界度量，即通过将参数向量按其相关性划分为多个子集（例如，可将模型的每一层视为一个子集），计算每个子集对应的 $\theta_c(\mathcal{V}_t^i)$ 并将其相加，可以显著减少所需的内存大小。极限情形下，可将参数向量 W 中的每一维均划分为单独子集，此时计算 θ_c 所需的内存大小等同于计算 θ_v，同时仍然比后者严格更紧。

2.4.2 随机梯度下降算法

与 SGLD 算法不同，SGD 算法在每步迭代中不涉及随机噪声，即 $\xi_t=0$。这使得以上用于推导 SGLD 泛化上界的策略不再可用，为 SGD 算法的泛化分析带来了额外

挑战。为此，Neu 等[34]引入了一种辅助优化轨迹 $\widetilde{W}_t$，在其中加入虚拟高斯噪声以模拟 SGLD 优化过程，最后建立这两种过程（W_t 和 $\widetilde{W}_t$）的泛化关联。令

$$\widetilde{W}_0 = W_0, \tag{2-29}$$

$$\widetilde{W}_t = \widetilde{W}_{t-1} - \eta_t g(W_{t-1}, B_t) + \widetilde{\xi}_t, \quad \text{若 } t > 0, \tag{2-30}$$

其中 $\widetilde{\xi}_t \sim N(0, \sigma_t^2 I)$ 为随机高斯向量。显然，以上辅助轨迹满足 $\widetilde{W}_t = W_t + \Delta_t$，其中 $\Delta_t = \sum_{i=1}^{t} \widetilde{\xi}_i$。基于这种思想，Wang 等[141]建立了如下的 SGD 泛化上界：

引理 2.15 ([141]，定理 2) 假设 $L(W_T) \le \mathbb{E}_{\Delta_t}[L(W_T + \Delta_t)]$，且 ℓ 二阶可微。则对于任意 $\sigma_1, \cdots, \sigma_T > 0$，有

$$|\overline{\text{gen}}| \le \frac{1}{2}\sum_{t=1}^{T} \sigma_t^2 \mathbb{E}_{W_T}[\mathbb{H}(W_T)] + |\overline{\text{gen}}(\widetilde{W}_T; \mathbf{Z})|, \tag{2-31}$$

$$I\big(\widetilde{W}_T; \mathbf{Z}\big) \le \sum_{t=1}^{T} \frac{d}{2} \log\left(\frac{\eta_t^2 V_t}{d\sigma_t^2} + 1\right), \tag{2-32}$$

其中 $\mathbb{H}(W_T) = \mathbb{E}_Z[\text{tr}(H_{W_T}(Z))]$，$H_W(z)$ 为损失函数 $\ell(W, z)$ 关于 W 的 Hessian 矩阵。

类似地，可通过引入核化 Rényi 熵以缓解其对于模型维度 d 的强依赖性。

定理 2.16 在引理 2.15 的相同条件下：

$$I_1\big(\widetilde{W}_T; \mathbf{Z}\big) \le \sum_{t=1}^{T} \left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t + I\right| + E^{\kappa}_{\widetilde{W}_t|\widetilde{W}_{t-1}}\right), \tag{2-33}$$

$$I\big(\widetilde{W}_T; \mathbf{Z}\big) \le \sum_{t=1}^{T} \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t + I\right|. \tag{2-34}$$

上述定理基于核化 Rényi 熵框架建立了 SGD 算法的泛化误差上界，其中式 (2-34) 对应于 $\sigma_\kappa \to 0$ 的极限情况。类似地，上述结果通过引入梯度协方差得到了更紧的上界估计结果。此外，还可进一步引入梯度的高阶矩以求进一步改进。但考虑到这会带来显著的计算负荷，此类改进在实际场景中的应用将较为受限。

2.5 矩阵 Rényi 熵的快速近似算法

本节从数值线性代数角度构建了矩阵 Rényi 熵的快速近似算法。上述分析引入的核化 Rényi 熵可视为矩阵 Rényi 熵在无限样本情形下的延伸，故而此类近似算法同样适用于核化 Rényi 熵。矩阵 Rényi 熵的朴素算法需要计算 $n \times n$ 矩阵的特征值分解，其计算复杂度为 $O(n^3)$，无法应用于大规模数据样本的信息熵计算。根据定义 2.1，计算矩阵 Rényi 熵的关键在于计算 A^α 的迹，故可考虑结合矩阵迹估计方法设计近似算法。经典矩阵迹估计方法包括高斯迹估计子和 Hutchinson 估计子[187]，其可对于给定正定矩阵

A 生成 $\mathrm{tr}(A)$ 的无偏估计。近期，Meyer 等[188]进一步改进了 Hutchinson 算法的收敛阶，显著降低了估计误差的方差，称为 Hutch++ 算法：

算法 2-1 Hutch++ algorithm for implicit matrix trace estimation[188]

Input: Kernel matrix $A \in \mathbb{R}^{n\times n}$, number of random vectors s ($s \ll n$), positive matrix function $f(A)$.
Output: Approximation to $\mathrm{tr}(f(A))$.
1 Sample $S \in \mathbb{R}^{n\times \frac{s}{4}}, G \in \mathbb{R}^{n\times \frac{s}{2}}$ from standard Gaussian distribution;
2 Compute an orthonormal basis $Q \in \mathbb{R}^{n\times \frac{s}{4}}$ for the span of AS via QR decomposition;
3 **return** $Z = \mathrm{tr}\left(Q^\top f(A)Q\right) + \frac{2}{s}\mathrm{tr}\left(G^\top(I-QQ^\top)f(A)(I-QQ^\top)G\right)$.

当取 $f(A) = A^\alpha$ 时，上述算法即可以高概率返回矩阵 Rényi 熵 log 函数内部分的 $(1\pm\varepsilon)$ 数值近似结果，将熵估计问题转化为矩阵-向量乘法运算，同时将整体复杂度降低到 $O(n^2 s)$，相较于基于特征值的朴素 $O(n^3)$ 算法大大降低了计算资源消耗。

2.5.1 整数阶近似算法

当 $\alpha \in \mathbb{N}$ 时，对于任意实值向量 g，$A^\alpha \cdot g$ 的结果均可通过计算 α 次 A 与向量的乘法获得。此观察结果直接引出了如下的整数阶矩阵 Rényi 熵近似算法：

算法 2-2 Integer order matrix-based Rényi's entropy estimation

Input: Kernel matrix $A \in \mathbb{R}^{n\times n}$, number of random vectors s, integer order $\alpha \geq 2$.
Output: Approximation to $S_\alpha(A)$.
1 Run Hutch++ with $f(A) = A^\alpha$ and s random vectors;
2 **return** $\widetilde{S}_\alpha(A) = \frac{1}{1-\alpha}\log(Hutch{+}{+}(A^\alpha))$.

上述算法的计算复杂度仅为 $O(\alpha s n^2)$。以下定理建立了该算法的近似精度保证：

定理 2.17 设 $\widetilde{S}_\alpha(A)$ 为算法 2-2 取

$$s = O\left(\frac{1}{\varepsilon}\sqrt{\log\left(\frac{1}{\delta}\right)} + \log\left(\frac{1}{\delta}\right)\right), \tag{2-35}$$

时的近似结果，则在置信度 $1-\delta$ 下，有

$$\left|\widetilde{S}_\alpha(A) - S_\alpha(A)\right| \leq \varepsilon \cdot S_\alpha(A). \tag{2-36}$$

2.5.2 非整数阶方法

对于非整数阶 α 情形，$A^\alpha \cdot g$ 的值将不存在简单计算方法，故考虑引入 A^α 的多项式近似，通过依次计算 $A\cdot g$、$A^2\cdot g$、$\cdots$ 以获得 $A^\alpha \cdot g$ 的数值近似结果。Chebyshev 级

数是一类常用的多项式近似方法，给定解析函数 $f:[-1,1]\mapsto\mathbb{R}$，其对应的 Chebyshev 级数定义为：

$$f(x)=\frac{c_0}{2}+\sum_{k=1}^{\infty}c_kT_k(x),\qquad x\in[-1,1] \tag{2-37}$$

其中 $T_0(x)=1$，$T_1(x)=x$，且当 $k\geq 1$ 定义 $T_{k+1}(x)=2xT_k(x)-T_{k-1}(x)$。当取上式的前有限 m 项时，其系数可通过以下公式计算：

$$c_k=\frac{2}{m+1}\sum_{i=0}^{m}f(x_i)T_k(x_i), \tag{2-38}$$

其中 $x_i=\cos\left(\pi(i+1/2)/(m+1)\right)$。通过结合线性映射 $g:[-1,1]\to[u,v]$，Chebyshev 级数可用于逼近任意 $\lambda\in[u,v]$ 的幂函数 $f(\lambda)=\lambda^{\alpha}$，其中 $\widehat{T}_k=T_k\circ g^{-1}$，如以下算法所示：

算法 2-3 Non-integer order matrix-based Rényi's entropy estimation via Chebyshev series

Input: Kernel matrix $A\in\mathbb{R}^{n\times n}$, number of random vectors s, fractional order α, polynomial order m, eigenvalue lower & upper bounds u, v.

Output: Approximation to $S_\alpha(A)$.

1 Run Hutch++ with $f(A)=c_0/2+\sum_{k=1}^{m}c_k\widehat{T}_k(A)$ and s random vectors;

2 **return** $\widetilde{S}_\alpha(A)=\frac{1}{1-\alpha}\log(Hutch{+}{+}(A^{\alpha}))$.

上述算法的计算复杂度为 $O(smn^2)$。类似地，可建立该算法的近似精度保证：

定理 2.18 设 $\widetilde{S}_\alpha(A)$ 为算法 2-3 取

$$s=O\left(\frac{1}{\varepsilon|\alpha-1|}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right), \tag{2-39}$$

$$m=O\left(\sqrt{\kappa}\log\left(\frac{\kappa}{\varepsilon|\alpha-1|}\right)\right), \tag{2-40}$$

时的近似结果，其中 $\kappa=v/u$ 是 A 的条件数，则在置信度 $1-\delta$ 下，有

$$\left|\widetilde{S}_\alpha(A)-S_\alpha(A)\right|\leq\varepsilon\cdot S_\alpha(A). \tag{2-41}$$

以上分析仅适用于 $u>0$ 的情形，即核矩阵 A 须为满秩矩阵。在部分核函数选择（如多项式核）下，核矩阵可能出现不满秩情况，导致上述误差上界不再适用。为此，进一步证明如下定理：

定理 2.19 设 $\widetilde{S}_\alpha(A)$ 为算法 2-3 取

$$s=O\left(\frac{1}{\varepsilon|\alpha-1|}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right), \tag{2-42}$$

$$m = O\left((vn)^{\frac{1}{2\min(1,\alpha)}} \sqrt[2\alpha]{\frac{1}{\varepsilon|\alpha-1|}}\right), \tag{2-43}$$

时的近似结果，则在置信度 $1-\delta$ 下，有

$$\left|\widetilde{S}_\alpha(A) - S_\alpha(A)\right| \leq \varepsilon \cdot S_\alpha(A). \tag{2-44}$$

当 $u=0$ 时，由于幂函数 $f(x)=x^\alpha$ 在 $x=0$ 处存在奇点，上述定理无法重现定理 2.18 在 $u>0$ 情形下的对数收敛率，仅能达到多项式级的收敛速度。

2.5.3 近似复杂度下界

以上建立了矩阵 Rényi 熵的近似算法，评估了其相关理论性质，并引出了后续问题：以上定理所建立的 $O(1/\varepsilon)$、$O(\sqrt{\kappa})$ 或 $O(\sqrt[2\alpha]{1/\varepsilon})$ 收敛阶是否具有进一步改进空间，即是否达到了理论最优收敛阶。本小节将证明以上收敛阶的近似最优性，即其与理论最优阶的差距不超过一个对数系数。首先，本小节基于随机隐式迹估计的复杂度下界[188]，通过复杂度归约建立了向量乘法数量 s 的下界：

定理 2.20 对于任意仅通过矩阵向量乘法 $Ar_1,\cdots,Ar_m$ 访问 $n\times n$ 归一化核矩阵 A 的算法，其中 $r_1,\cdots,r_m$ 为算法自适应选择的随机向量，在有限精度计算模型下，至少需要

$$s = \Omega\left(\frac{1}{\varepsilon|\alpha-1|\log n \log(1/\varepsilon|\alpha-1|\log n)}\right), \tag{2-45}$$

次矩阵向量乘法查询以获得估计值 Z，使得对任意 $\alpha>0$，以至少 $\frac{2}{3}$ 的概率，有

$$|Z - S_\alpha(A)| \leq \varepsilon \cdot S_\alpha(A). \tag{2-46}$$

基于多项式最优一致逼近误差理论，可进一步建立多项式近似阶数 m 的下界：

定理 2.21 存在正值递减函数 $\varepsilon_0:\mathbb{R}^+\to\mathbb{R}^+$，使得对于任意 $0<u<v<1$ 和 $0<\varepsilon<\varepsilon_0(v/u)$，任意多项式 p_m 至少需要

$$m = \Omega\left(\sqrt{\kappa}\log\left(\frac{1}{\kappa\varepsilon|\alpha-1|\log n}\right)\right), \tag{2-47}$$

阶以实现对于任意特征值处于 $[u,v]$ 范围内且满足 $\mathrm{tr}(A)\in[1,2]$ 的核矩阵 A，有

$$\left|\frac{1}{1-\alpha}\log\left(\mathrm{tr}(p_m(A))\right) - S_\alpha(A)\right| \leq \varepsilon \cdot S_\alpha(A). \tag{2-48}$$

定理 2.22 对于任意 $v>0$ 和足够小的 ε, 任意多项式 p_m 至少需要

$$m = \Omega\left(\sqrt[2\alpha]{\frac{1}{\varepsilon|\alpha-1|\log n}}\right), \tag{2-49}$$

阶以实现对于任意特征值处于 $[0, v]$ 范围内的核矩阵 A，有

$$\left|\frac{1}{1-\alpha}\log\left(\mathrm{tr}(p_m(A))\right) - S_\alpha(A)\right| \leq \varepsilon \cdot S_\alpha(A). \tag{2-50}$$

可注意到，定理 2.21 中的下界 $s = \Omega(1/\varepsilon)$ 与此前上界定理中的收敛阶是一致的。此外，上述多项式阶理论下界也分别与定理 2.18 与定理 2.19 中的上界收敛阶吻合。这些理论结果共同证明了算法 2-3 复杂度的理论最优性。

2.6 实验分析

本节将上述面向 SGLD/SGD 等随机迭代学习算法的泛化上界进行可视化，以验证这些结果相较于已有工作在紧致性以及可计算性上的改进。简便起见，以下使用恒定的学习率 $\eta_t = \eta$ 与高斯噪声 $\sigma_t = \sigma$。通过选择合适的 σ_κ 值，期望偏差 E_X^κ 将趋于 0，故而可在计算中忽略不计。定理 2.10 中的次高斯常数可通过收集每个批次中的损失值，并计算 $\sigma = \frac{1}{2}[\max_t \ell(W_{t-1}, B_t) - \min_t \ell(W_{t-1}, B_t)]$ 以获得其上界估计。上述互信息上界中的 $\mathcal{V}_t$ 与 V_t 可使用 BackPack 库以从每个批次中获得其经验估计。定理 2.16 中的 $\mathbb{H}(W_T)$ 则可使用 PyHessian 库计算相关 Hessian 矩阵。对于每项实验，重复独立训练 100 次以获得 W_t 和 B_t 的独立同分布样本，并依此构建式 (2-10) 中的核矩阵 K，用于计算相关信息论泛化上界中的核化 Rényi 互信息。

2.6.1 矩阵 Rényi 熵近似性能

为了评估矩阵 Rényi 熵相关近似算法，通过混合高斯分布 $\frac{1}{2}N(-1, I_d) + \frac{1}{2}N(1, I_d)$ 生成模拟数据，其中 $n = 5,000$，$d = 10$，I_d 是 d 维单位矩阵。相应核矩阵大小即为 $5,000 \times 5,000$。以下使用高斯核函数 $\varphi(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/2\sigma^2)$ 计算矩阵 Rényi 熵，其中 $\sigma = 1$。对于每项实验，以下将报告 100 次试验后近似结果的平均相对误差（MRE）及其标准差（SD）。基准值 $S_\alpha(A)$ 可通过 $O(n^3)$ 的特征值分解方法获得。实验环境为 Intel i7-10700 (2.90GHz) CPU，内存 64G。

首先评估算法 2-2 对整数阶矩阵 Rényi 熵的近似性能。以下分别给出了 $\alpha \in \{2, 3, 5, 8\}$ 时平均相对误差随向量数量 s 的变化曲线，其中 s 取值范围为 10 到 150，如图 2-2 所示。阴影区域表示对应的标准差。容易观察到在 log 坐标轴尺度下，s 与 MRE 呈线性关系，与定理 2.17 的误差上界相吻合。特别地，当 $\alpha = 2$ 时，仅需 $s = 10$ 个随机向量即可实现 0.1% 的平均相对误差，约为 1.2 秒运行时间。相比之下，朴素的特征值方法需要 27 秒以计算完整的特征值分解。

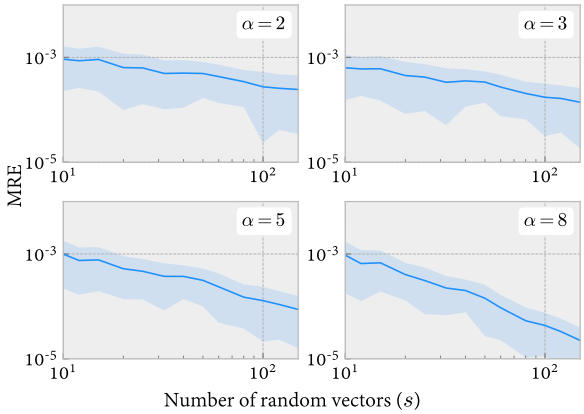


图 2-2 整数阶矩阵 Rényi 熵近似中，平均相对误差随向量数量 s 的变化曲线

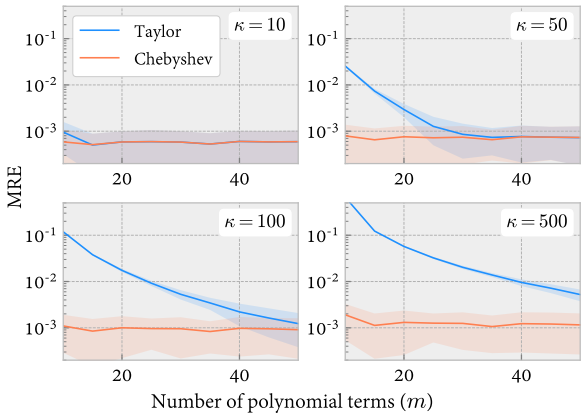


图 2-3 非整数阶矩阵 Rényi 熵近似中，平均相对误差随条件数 κ 的变化曲线

下面探讨不同条件数 κ 对于多项式近似精度的影响。设 $\alpha = 1.5$，$s = 100$，m 的取值范围为 10 到 50。核矩阵的特征谱可通过调整高斯核函数中的宽度参数 σ 控制。图 2-3 对比了 Chebyshev 级数相较于 Taylor 级数的近似性能。可见对于较大的 κ，Taylor 级数所需的多项式阶数多于 Chebyshev 级数。

2.6.2 模拟数据上的泛化上界可视化

对于模拟数据，考虑简单的线性回归问题

$$y = Wx + \epsilon, \tag{2-51}$$

其中 x 为 10 维输入向量，y 为目标值，W 为线性回归系数，ϵ 为零均值的随机噪声。所训练的神经网络为 3 层 MLP 网络，其中隐层宽度为 10。对于每次独立的训练过程，使用相同策略生成大小为 $n = 100$ 的训练数据集。现有理论泛化上界与真实泛化误差的比较如图 2-4 左侧所示，其中不同的关键信息度量代表了不同的泛化上界估计：IWZ 对应 $I_1(W_T; \mathbf{Z})$，IWB|W 对应 $I_1(W_t; B_t|W_{t-1})$ 求和。可见 IWZ 曲线始终高于 IWB|W，且两

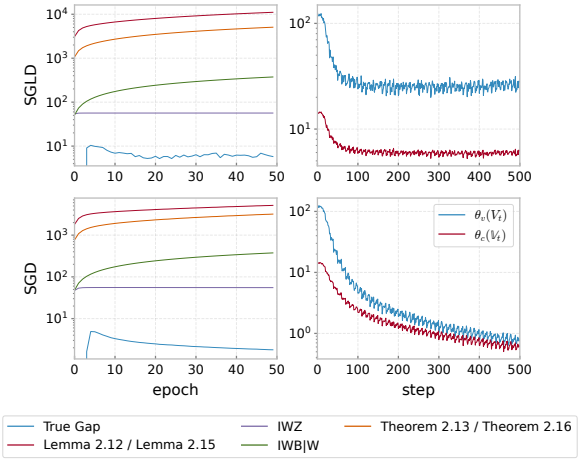


图 2-4 模拟数据上的泛化上界对比

者均始终位于 SGLD（或 SGD）的泛化理论上界与真实泛化误差曲线之间。这表明上述可视化结果成功反映了其关键信息量 $I(W_t;\mathbf{Z})$ 和 $I(W_t;B_t|W_{t-1})$ 的实际变化曲线。此外，该结果还表明：

（1） IWZ 与真实泛化误差之间的差距表明，损失函数 $\ell(w,Z)$ 相对于 Z 的次高斯常数 σ 存在过度估计问题。正常而言，经过良好训练的神经网络损失值应低于随机初始化网络的损失值，因此常数 σ 理应随训练的推进而逐步减小。这一观察可用于相关上界的进一步改进。

（2） 如图 2-5 所示，IWZ 在训练过程中迅速达到峰值（或开始下降），其变化趋势与真实泛化误差曲线吻合。而 IWB|W 则始终上升，因为 $I(W_t;B_t|W_{t-1})$ 始终为正。这一观察表明，尽管模型总是从新批次中学习一些新知识（即 $I(W_t;B_t|W_{t-1})\geq 0$），但 W 关于数据集 $\mathbf{Z}$ 包含的总信息（即 $I(W_t;\mathbf{Z})$）迅速达到上限后不再增加，这表明网络实际遗忘了部分先前学习的知识以获得更强的泛化能力，这种现象也被称为神经网络的隐式正则化。然而，现有信息论泛化上界无法刻画这种“遗忘”能力，导致 IWZ 与 IWB|W 间的差距在训练过程推进时逐渐增大。

（3） IWB|W 与本章的改进上界间仍存在差距，这表明即使考虑了梯度不同维度间的相关性，当前上界仍远未达到最优。这一观察也与近期工作[189-190]吻合，即 SGD 生成的随机梯度向量拥有重尾分布，故而使用高斯假设将显著高估其熵值。另一种猜想是神经网络的隐式正则化机制[191-192]，这将导致网络权重所捕获的信息远低于其理论容量。

图 2-4 右侧则提供了梯度协方差上界 θ_c 和梯度方差上界 θ_v 的直观对比。可见 $\theta_v(V_t)$ 始终大于 $\theta_c(\mathcal{V}_t)$，特别是在训练过程的开始阶段。这一观察结果验证了定理 2.14。

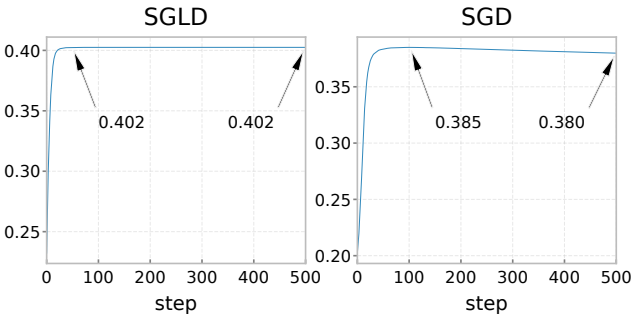


图 2-5 $I_1(W_T;\mathbf{Z})$ 在模拟数据上的变化曲线

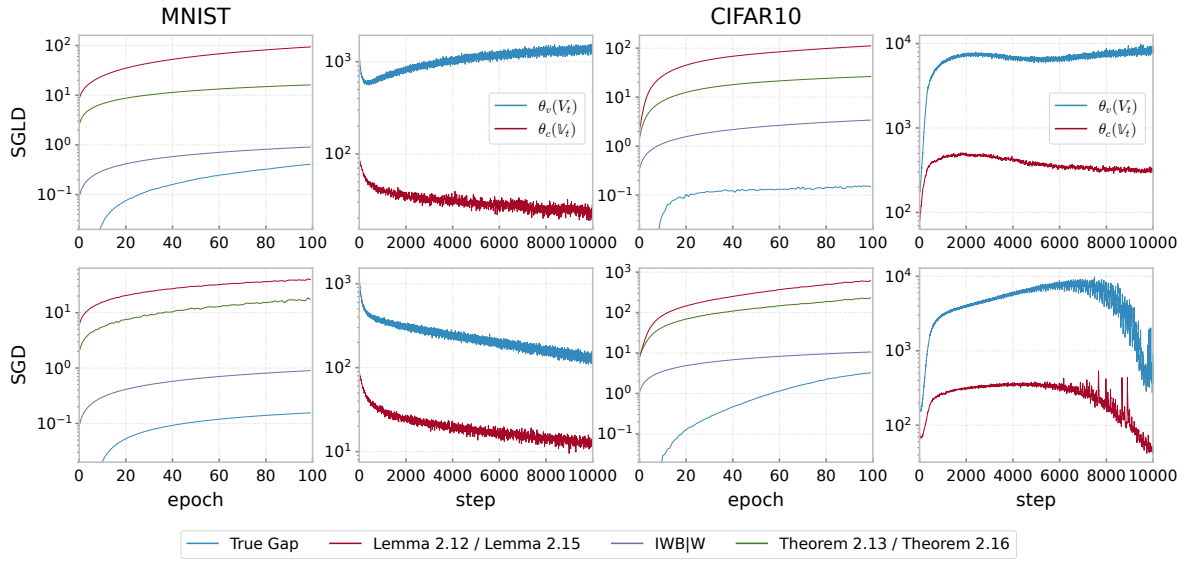


图 2-6 真实数据上的泛化上界对比

2.6.3 真实数据上的泛化上界可视化

以下将在真实数据集上可视化相关泛化上界，以展示核化 Rényi 熵的可扩展性。遵循 Wang 等[138]的实验设置，在 MNIST 数据集上训练一个更宽的 MLP 网络，并在 CIFAR10 数据集上训练一个 4 层 CNN 网络，对比结果如图 2-6 所示。可见，IWB|W 曲线始终落在相邻界之间的正确区间内，并准确反映了梯度协方差上界的上升趋势。梯度方差上界仍显著高于 IWB|W，相比之下，基于梯度协方差的改进上界一定程度上填补了其间隙。此外，SGLD 算法的 θ_c 曲线在训练后期迅速停止上升并开始下降，这预示着网络容量接近饱和；而 θ_v 曲线则始终随着训练过程推进而上升。这一观察进一步验证了本章中改进泛化上界的紧致性。

以下将进行随机标签实验，以展示在不同标签噪声水平下改进泛化上界的紧致性。基本设置与上述实验保持相同，同时依据超参数 ρ 以一定概率随机替换训练标签为噪声标签。如图 2-7 所示，更高概率的标签噪声将导向更高的泛化误差。虽然 Wang 等[141]的泛

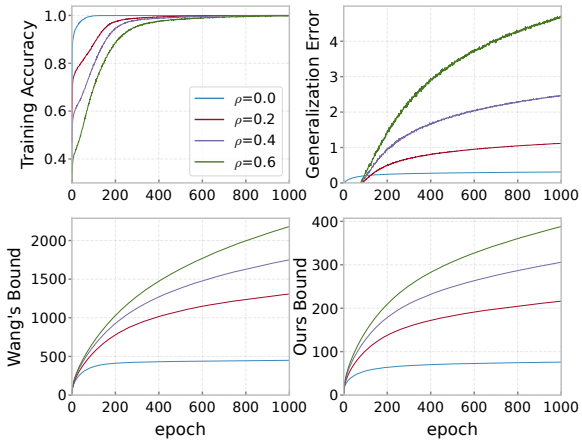


图 2-7 MNIST 上的随机标签实验

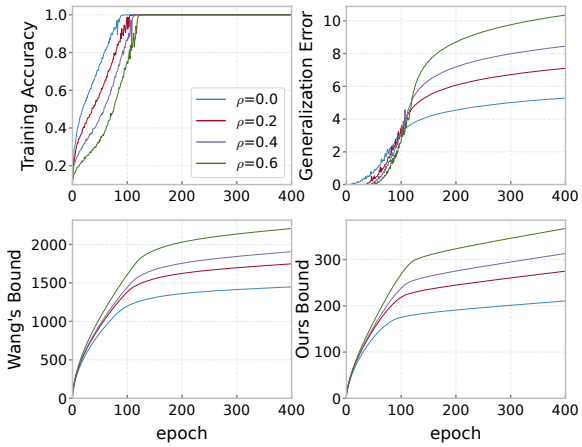


图 2-8 CIFAR10 上的随机标签实验

化上界与本章所构建的上界均准确刻画了真实泛化误差的变化趋势，但本章中的上界估计显著更紧，更能够反映泛化误差的实际值。图 2-8 展示了随机标签实验在 CIFAR10 数据集上的结果，图 2-9 则展示了不同隐层宽度下 MLP 网络的泛化性能。可见，本章中的改进泛化上界相较于现有工作平均有 5 倍以上的紧致度提升。

2.7 本章小结

本章解决了传统 Shannon 信息度量在实际应用中难以量化计算的问题，这种问题使得通过假设互信息度量构建的信息论泛化误差上界无法计算，从而难以发现目前理论推导的瓶颈所在，难以针对目前泛化上界的过度估计问题做出有效改进。为此，本章提出了一种新型信息度量方法，称为核化 Rényi 熵，它能够通过有限采样数据直接近似计算而不受随机变量维度影响，且能够兼容现有的泛化分析框架。在此之上，成功基于核化 Rényi 熵推广了现有面向随机迭代学习算法的信息论泛化上界，并分析了多种后

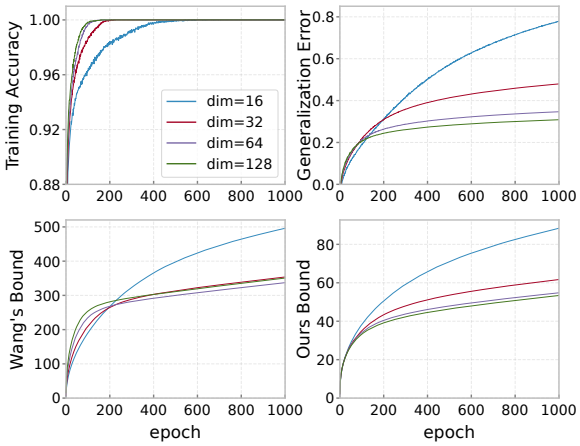


图 2-9 不同隐层宽度对于 MLP 泛化能力的影响

续改进策略。通过引入梯度轨迹协方差，获得了 SGLD 与 SGD 算法相较于目前结果更紧的上界。进一步地，针对矩阵 Rényi 熵计算效率低下的问题，基于矩阵迹估计与多项式近似方法构建了相关快速近似算法，并通过理论分析证明了它们的最优性，为相关泛化上界的可计算性提供切实保障。

本章有关泛化理论分析的研究工作发表于人工智能顶级会议、CCF 推荐 A 类学术会议 International Joint Conference on Artificial Intelligence，论文题目为“Understanding the Generalization Ability of Deep Learning Algorithms: A Kernelized Rényi’s Entropy Perspective”；有关矩阵熵快速近似算法的研究工作发表于机器学习理论顶级期刊、CCF 推荐 A 类学术期刊 IEEE Transactions on Information Theory，论文题目为“Optimal Randomized Approximations for Matrix-based Rényi’s Entropy”。

3 损失熵引导的高概率信息论泛化误差上界

3.1 引言

近年来，基于信息论度量分析相关学习算法泛化性质的方法与技术逐渐引起了国内外学者的广泛关注[28,193]。其核心概念涉及量化模型权重中所储存的有关训练数据集的信息量，其可作为深度学习模型过拟合的自然指标。此类信息论泛化上界具备多种优势：相关信息论度量不仅能够同时刻画数据分布与学习算法的相关性质，同时其基本假设相较于其他泛化分析技术（如一致稳定性[174,177-178] 或模型压缩[181-182]）而言更为宽松。基于信息论的泛化分析已成为刻画随机迭代学习算法泛化能力的有力工具[32-34,54-55,141,194]。

然而，相关研究显示此类泛化上界存在严重过度估计问题：在实践中，深度神经网络往往能够在记忆大量训练数据的同时仍具备良好的泛化能力[195]。此外，由于 Shannon 熵理论的根本性限制，构建相关泛化上界所依赖的信息论度量在面对高维随机变量（如大规模神经网络）时往往难以量化计算。Steinke 等[32]首次提出了基于超样本技术的泛化分析方法，通过从称为超样本集的大规模数据样本中划分训练集与测试集，随后基于训练集选择变量与其他多种度量（例如网络预测值[58]、损失对[51]或损失差异[130]）之间的互信息以界定泛化误差所在范围，取得了可计算性方面的重要进展。值得注意的是，此类上界不仅通过在关键信息度量中引入低维随机变量而提高了可计算性，且在面对大型神经网络模型时往往能够提供更紧的上界估计。

虽然上述工作已在改进信息论相关泛化上界的可计算性方面取得了可喜的成果，但此类上界结果多局限于期望情形下的泛化误差刻画，缺乏可计算的高概率泛化上界研究。此外，此类上界多依赖于损失函数的有界性假设，导致其在面对长尾损失分布（如交叉熵损失）时不再适用。本章构建了多种仅包含一维随机变量的信息论泛化上界，并展示了此类上界足以准确刻画相关学习算法的实际泛化能力。本章提出的泛化上界不仅从根本上解决了由高维随机变量带来的可计算性挑战，且在现有结果的基础之上进一步放宽了损失函数的有界性假设，从而适用于任意无界损失函数。

具体而言，本章首先展示了 Kawaguchi 等[46]工作中面向信息瓶颈所构建的泛化上界的一种改进方法：通过将模型输入数据与中间表示之间的互信息替换为损失熵，能够有效解决相关互信息度量引起的潜在无界问题，获得显著更紧的泛化估计结果。此外，仅包含一维随机变量的损失熵显著改善了相关上界的可计算性。基于损失熵度量，本章推导了多种数据无关情形下的高概率泛化误差上界，分别使用了非条件（定理 3.1）与条件（定理 3.2）信息度量。值得注意的是，此类上界为最小化误差熵 (Minumum Error Entropy, MEE) 准则[196] 提供了新的见解：首次展示了真实泛化误差与 $\sqrt{H(E^w)/n}$ 之间的比例关系，提出了基于 MEE 准则的信息论泛化上界，其中 E^w 为预测误差，n 为训练

集的大小。进一步地，本章基于损失熵简化了数据依赖情形下，现有结果中依赖于高维信息度量的泛化误差上界。此类改进上界同样仅依赖于一维随机变量（损失或损失差异）的相关信息度量，从而不仅显著提高了可计算性，并且在留一法（定理 3.3）与超样本（定理 3.4）设定下放宽了损失函数的严格有界假设。此外，本章后续通过引入广义加权泛化误差构建了具备快速收敛率的泛化上界（定理 3.5），此类上界可在经验风险接近 0 时，将收敛速率从 $1/\sqrt{n}$ 提升到 $1/n$。通过进一步结合阈值策略（定理 3.6），此类上界克服了目前快速收敛率泛化结果在面对长尾损失分布时的关键限制。最后，这些理论结果在多个模拟与真实数据集上得到了验证，相关实验证实了真实泛化误差值与本章提出的损失熵泛化度量之间的强相关性。此外，本章的数据依赖泛化上界准确刻画了多项深度学习任务中的泛化误差变化曲线，相较于目前最优的高概率信息论泛化上界[51]在紧致性与可计算性方面均有显著提升。

总体而言，本章的主要贡献包括：（1）提出新型低维信息论泛化度量：损失熵。该度量仅包含一维随机变量，可通过核密度估计、分箱方法 (Binning Method) 等直接近似计算，为真实神经网络的泛化能力提供数值估计结果。（2）结合基于典型子集及超样本集的泛化分析技术，构建了基于损失熵的高概率泛化误差上界估计，可精确刻画随机算法的实际泛化性能，为复杂随机学习算法的泛化分析提供了新思路和新方法。（3）在多项深度学习任务上验证了相关理论结果，发掘损失熵度量与泛化误差间的强相关性，相较于现有泛化理论结果显著提升了紧致性与可计算性。

3.2 基本概念与问题设定

为方便后续讨论，假设后续理论分析中所有涉及的损失变量均为离散型随机变量，且其取值空间基数有限。后续分析将讨论一种针对连续型损失函数的离散化方法，使得相关理论分析同样适用于此类损失函数。对于给定数据样本 $Z \sim \mu$，设 $L^w = \ell(w, Z)$ 为其对应的损失值，且设 $L_i^w = \ell(w, Z_i)$ 为训练数据对应的损失值。此外，引入 $b^w = \sup_{z\in\mathcal{Z}} \ell(w,z)$ 表示给定假设 $w \in \mathcal{W}$ 时的最大可达损失，以及 $B^{w,\mathbf{Z}} = \sup_{i\in[1,n]} L_i^w$ 表示给定假设在数据集 $\mathbf{Z}$ 上的最大样本损失。

留一法（LOO）设定在近期由 Rammal 等[197]引入用于信息论泛化分析。设 $\widetilde{\mathbf{Z}}_l = \{Z_i\}_{i=1}^{n+1} \in \mathcal{Z}^{n+1}$ 为包含 $n+1$ 个独立同分布样本的数据集。$U \sim \mathrm{Unif}([1, n+1])$ 为一个离散均匀随机变量，代表从 $\widetilde{\mathbf{Z}}_l$ 中选择的单个测试样本。随后，构建训练集 $\mathbf{Z}_l = \widetilde{\mathbf{Z}}_l \setminus Z_U$ 与测试集 $\overline{\mathbf{Z}}_l = \{Z_U\}$。此外，定义 $R^w = \{L_i^w\}_{i=1}^{n+1} \in \mathcal{R}^w$ 为所有数据样本对应损失的集合，假设 W 的泛化能力则通过留一法验证误差 $\overline{\mathrm{gen}}(W, \widetilde{\mathbf{Z}}_l, U) = L_{\overline{\mathbf{Z}}_l}(W) - L_{\mathbf{Z}_l}(W)$ 衡量。

基于**超样本**技术的泛化分析框架最初由 Steinke 等[32]引入。设 $\widetilde{\mathbf{Z}}_s = \{Z_{i,0}, Z_{i,1}\}_{i=1}^n \in \mathcal{Z}^{n\times 2}$ 为包含 $n\times 2$ 个独立同分布样本的数据集。$\widetilde{U} = \{\widetilde{U}_i\}_{i=1}^n \sim \mathrm{Unif}(\{0,1\}^n)$ 为用于区分训练与测试样本的 n 个随机 $\{0,1\}$ 变量，其中 $\widetilde{U}_i = 0$ 表示 $Z_{i,0}$ 将用于模型训练而 $Z_{i,1}$ 将

用于测试，反之亦然。随后，构建训练集 $\mathbf{Z}_s = \{Z_{i,\widetilde{U}_i}\}_{i=1}^n$ 与测试集 $\overline{\mathbf{Z}}_s = \{Z_{i,1-\widetilde{U}_i}\}_{i=1}^n$。此外，定义 $\widetilde{R}^w = \{L_{i,0}^w, L_{i,1}^w\}_{i=1}^n \in \widetilde{\mathcal{R}}^w$ 为所有数据样本对应损失的集合，以及 $\widetilde{R}_\Delta^w = \{\Delta L_i^w\}_{i=1}^n \in \widetilde{\mathcal{R}}_\Delta^w$ 为损失差异的集合，其中 $\Delta L_i^w = L_{i,1}^w - L_{i,0}^w$。假设 W 的泛化能力则通过验证误差 $\overline{\mathrm{gen}}(W, \widetilde{\mathbf{Z}}_s, \widetilde{U}) = L_{\overline{\mathbf{Z}}_s}(W) - L_{\mathbf{Z}_s}(W)$ 衡量。

3.3 主要定理

本节通过构建信息论视角下的高概率泛化误差上界，探讨了损失熵与真实泛化误差的关联。以下首先考虑了假设固定且独立于训练数据集的学习场景，在其中改进了 Kawaguchi 等[46]工作中面向信息瓶颈方法的泛化上界。随后，在留一法与超样本问题设定下构建了全新的数据依赖泛化上界。最后，基于广义加权验证误差构建了具备快速收敛阶的泛化上界[116,198]。

3.3.1 数据无关的泛化上界

本小节通过数据无关场景下的泛化分析，表明在假设网络参数 w 固定且独立于训练数据集 $\mathbf{Z}$ 的情形下，损失熵与模型的泛化能力之间存在强相关性，因而最小化损失熵准则有利于增强深度学习模型的泛化能力。具体而言，给定任意 $w \in \mathcal{W}$，随机样本 $Z \sim \mu$ 且设 $L^w = \ell(w, Z)$，则有以下高概率泛化误差上界：

定理 3.1 给定任意 $\gamma > 0$ 与 $\delta > 0$，则以置信度 $1-\delta$，有:

$$\overline{\mathrm{gen}}(w, \mathbf{Z}) \leq C_1^w \sqrt{\frac{H(L^w) + C_2^w}{n}} + \frac{C_3^w}{\sqrt{n}}, \quad \text{其中} \begin{cases} C_1^w = 2b^w\sqrt{2} \\ C_2^w = c^w\sqrt{\frac{m\log(\sqrt{n}/\gamma)}{2}} + \log(2/\delta) \\ C_3^w = \gamma b^w + \frac{B^{w,\mathbf{Z}}\sqrt{\gamma}}{n^{1/4}}\sqrt{2\log(2/\delta)} \end{cases} \tag{3-1}$$

定理 3.2 给定任意 $\gamma > 0$ 与 $\delta > 0$，则以置信度 $1-\delta$，有:

$$\overline{\mathrm{gen}}(w, \mathbf{Z}) \leq \widetilde{C}_1^w \sqrt{\frac{H(L^w|Y) + \widetilde{C}_2^w}{n}} + \frac{\widetilde{C}_3^w}{\sqrt{n}}, \quad \text{其中} \begin{cases} \widetilde{C}_1^w = 2b^w\sqrt{2|\mathcal{Y}|} \\ \widetilde{C}_2^w = c^w\sqrt{\frac{m\log\left(\sqrt{n}/\gamma\right)}{2}} + \log(2|\mathcal{Y}|/\delta) \\ \widetilde{C}_3^w = \gamma b^w + \frac{B^{w,\mathbf{Z}}\sqrt{\gamma|\mathcal{Y}|}}{n^{1/4}}\sqrt{2\log(2|\mathcal{Y}|/\delta)} \end{cases} \tag{3-2}$$

在上述定理中，m 与 c^w 分别代表扰动变量的维度和敏感度（数学定义见附录 A.3），是数据分布 μ 中随机性的来源。容易验证，当 $n \to \infty$ 时，有 $C_1^w, C_2^w, C_3^w = \widetilde{O}(1)$，从而上述泛化误差上界拥有收敛阶 $1/\sqrt{n}$。类似结论同样适用于 $\widetilde{C}_1^w, \widetilde{C}_2^w, \widetilde{C}_3^w$。值得注意的是，由于采用了模型最后一层的输出（即损失）构建信息论度量，以上上界适用于任意可通过函数 $f: \mathcal{Z} \mapsto \mathbb{R}^+$ 表示的确定型神经网络结构与损失函数，而无需类似于 Kawaguchi

等[46]上界中所要求的网络中间表示层。

定理 3.1 与定理 3.2 间的对比强调了关于 $|\mathcal{Y}|$（标签空间的基数）复杂性的权衡：一方面，由于条件熵总是小于对应的非条件熵，可得 $H(L^w|Y) \leq H(L^w)$。另一方面，定理 3.2 中的上界估计关于 $\sqrt{|\mathcal{Y}|}$ 呈线性增长关系，从而在 $|\mathcal{Y}|$ 较大时（例如回归问题）将导致相关上界失效。当 $I(L^w;Y)$ 较大，即网络在不同类别上的性能差异较为显著时，推荐采用定理 3.2 以获得更紧的上界估计。否则，建议优先采用定理 3.1 中的上界。

相较于 Kawaguchi 等[46]工作中的数据无关泛化理论结果，定理 3.2 的一个显著区别在于使用 $H(L^w|Y)$ 替代了 $I(X;T^w|Y)$，其中 T^w 代表神经网络模型中编码器部分产生的中间特征表示。容易验证在给定 Y 时，$X \to T^w \to L^w$ 可构成条件 Markov 链，进一步结合数据处理不等式，可证明互信息 $I(L^w;X|Y) = H(L^w|Y) - H(L^w|X,Y)$ 相较于信息瓶颈的优化目标 $I(X;T^w|Y)$ 而言，可导向严格更紧的泛化误差上界。当相应模型架构为确定型（即模型前向运算过程中不涉及外部随机量，在现代深度神经网络结构如 MLP、CNN 等中十分常见）时，易证 $H(L^w|X,Y) = 0$，这意味着 $I(L^w;X|Y)$ 与 $H(L^w|Y)$ 等价。此外，在部分已知特殊情形（如当 X 的全部信息可以通过 T^w 完整恢复时）下，互信息 $I(X;T^w|Y)$ 可能出现无界情形[199]。与之相比，损失熵 $H(L^w)$ 和 $H(L^w|Y)$ 始终有界。

上述泛化理论上界带来的另一个改进涉及系数 C_1^w，其在 Kawaguchi 等[46]工作中的定义中包含 $\sum_{t\in\mathcal{T}_\varepsilon^w} \mathbb{P}^{1/2}(T^w = t)$，其中 $\mathcal{T}_\varepsilon^w \approx \mathcal{T}^w$ 为表示空间 $\mathcal{T}^w$ 的典型子集。此度量将随 $H_{1/2}(T^w) > H(T^w)$ 呈指数形式增长，从而需要对 T^w 概率分布的衰减率做出额外假设以保证 C_1^w 有界并保持该上界的收敛率为 $1/\sqrt{n}$。本小节改进了此定理证明中的相关技巧，通过在应用多项式集中不等式时采用留一法策略，克服了这一限制，得到了与 L^w 概率分布无关的常数 C_1^w 系数，且无需任何额外假设。

此外，以上数据无关场景下的泛化上界为理解最小化误差熵准则的泛化能力提供了全新见解。通常而言，对于回归或二分类任务，给定样本 $Z \sim \mu$ 的预测误差定义为 $E^w = Y - f(w,X)$ 网络输出与实际标签间的差值，其中 f 为网络的编码器部分。给定损失函数 ℓ，最终损失即可通过给定预测误差的确定型映射 $L^w = \ell(E^w)$ 计算（如均方损失 $\ell(x) = x^2$）。基于数据处理不等式，易证 $H(L^w) \leq H(E^w)$，这表明优化 $H(E^w)$ 可同时最小化定理 3.1 中的上界估计，从而验证了最小化误差熵准则能够增强深度学习模型的泛化能力。类似地，对于交叉熵等基于间隔的损失函数，最终损失可视为给定预测误差与实际标签的确定型映射 $L^w = \ell(E^w, Y)$，从而根据条件数据处理不等式易证 $H(L^w|Y) \leq H(E^w|Y)$，结合 $H(L^w|Y) \leq H(L^w)$ 即可得到相同结论。

最后，上述泛化误差上界解决了 Kawaguchi 等[46]结果中有关可计算性的重要问题。由于输入 X 与表示 T^w 均为高维随机变量，特别是对于现代大型神经网络而言，导致准确计算 $I(X;T^w|Y)$ 的值极其困难。这为使用此类泛化上界以评估深度学习模型的实际泛化能力带来了额外的困难。相比之下，本小节中的上界仅涉及 L^w 和 Y，其对于绝大多

数机器学习任务而言均为一维或离散随机变量，从而使得估计 $H(L^w)$ 或 $H(L^w|Y)$ 的值不仅在实际中可行，而且十分高效。

虽然在实际应用中，模型 w 并不总是独立于训练数据集 $\mathbf{Z}$，定理 3.1 与 3.2 在特定学习任务中仍然具有指导意义。其中包括预训练任务，其目标为在特定数据集上评估预训练模型的泛化能力。另一种相关任务是评估模型在验证数据集上的性能，本小节的上界为平均验证误差偏离总体风险的幅度与概率提供了有效保障。

3.3.2 数据依赖的泛化上界

本小节将以上泛化分析方法扩展到数据依赖的学习场景中，其中模型 W 将在训练期间学习数据集 $\mathbf{Z}$ 的相关信息，从而成为与 $\mathbf{Z}$ 相关的随机变量。为此，本小节在留一法与超样本设定下分析了相关深度学习模型与算法的泛化性能，并分别构建了两种学习场景下的高概率泛化误差上界。

定理 3.3 给定任意 $\lambda \in (0,1)$ 与 $\delta > 0$，则以置信度 $1-\delta$，有:

$$\overline{\mathrm{gen}}\left(W, \widetilde{\mathbf{Z}}_l, U\right) \leq C_1^W \sqrt{H_{1-\lambda}(R^W) + C_2^W}, \quad \text{其中} \begin{cases} C_1^W = \sqrt{2}\Sigma_{R^W} \\ C_2^W = \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \log\left(\frac{2}{\delta}\right) \end{cases} \tag{3-3}$$

其中，$\Sigma_{R^W} \in [0, B^{W,\widetilde{\mathbf{Z}}_l}]$ 为 $\frac{n+1}{n}(L_U^W - \overline{L}^W)$ 相对于 U 的次高斯常数。

定理 3.4 给定任意 $\lambda \in (0,1)$ 与 $\delta > 0$，则以置信度 $1-\delta$，有:

$$\overline{\mathrm{gen}}\left(W, \widetilde{\mathbf{Z}}_s, \widetilde{U}\right) \leq \widetilde{C}_1^W \sqrt{\frac{H_{1-\lambda}(\widetilde{R}_\Delta^W) + \widetilde{C}_2^W}{n}}, \quad \text{其中} \begin{cases} \widetilde{C}_1^W = \sqrt{\frac{2}{n}\sum_{i=1}^n \left(\Delta L_i^W\right)^2} \\ \widetilde{C}_2^W = \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \log\left(\frac{2}{\delta}\right) \end{cases} \tag{3-4}$$

其中，当 $n \to \infty$ 时容易验证 $\lim_{n\to\infty} \overline{\mathrm{gen}}(W, \widetilde{\mathbf{Z}}_s, \widetilde{U}) = \overline{\mathrm{gen}}(W, \mathbf{Z}_s)$，即在数据集足够大时，超样本设定下的验证误差可作为真实泛化误差的良好近似。类似地，当 $n \to \infty$ 时同样有 $\widetilde{C}_1^W, \widetilde{C}_2^W = \widetilde{O}(1)$。其中，参数 λ 隐含了样本损失的联合 Rényi 熵与上界置信度之间的权衡：$H_{1-\lambda}(X)$ 将随 λ 的增加而单调上升。值得注意的是，定理 3.4 拥有 $1/\sqrt{n}$ 的显式收敛阶，而对定理 3.3 则不成立。这是因为在留一法设定下，测试损失（即总体风险的替代度量）仅通过单个样本 Z_U 进行评估，从而导致验证误差的方差较高。类似现象也可在 Rammal 等[128,197]的结果中观察到。

定理 3.3 的主要优势来源于其在插值模式 (Interpolating Regime) 下的适用性，即模型总是能够达到 0 训练损失。在这种情形下，$H(R^W)$ 可进一步简化为 $H(L_U^W)$，即一个一维随机变量的熵度量。这一特性使得通过提供测试损失 L_U^W 的独立同分布样本，便可直接量化计算这一上界估计。类似地，在插值模式下，定理 3.4 中的 $H(\widetilde{R}_\Delta^W)$ 可简化为 $H(\{L_{i,1-\widetilde{U}_i}\}_{i=1}^n)$，即所有测试样本损失的联合熵。虽然这一信息度量仍然包含高维随机

变量从而不易于直接计算，但可进一步利用 Shannon 熵的次可加性以构建其替代上界：$H(\{L_{i,1-\widetilde{U}_i}\}_{i=1}^n) \leq \sum_{i=1}^n H(L_{i,1-\widetilde{U}_i})$。这一特性同样适用于当 $\lambda \approx 0$ 时的 Rényi 熵，从而确保了定理 3.3 与定理 3.4 的可计算性。

定理 3.4 进一步利用了训练与测试损失值之间的损失差异以推导严格更紧的泛化上界，其中损失差异可表示为 $\Delta L_i^W = L_{i,1}^W - L_{i,0}^W$。这一概念首次在 Wang 等[130]的工作中得到探索，其证明了通过数据处理不等式，可得互信息 $I(\Delta L_i^W, \widetilde{U}_i) \leq I(L_{i,1}^W, L_{i,0}^W; \widetilde{U}_i)$。此处，可结合 Shannon 熵的凹性将这一结论拓展至损失熵度量。具体而言，观察到 $\frac{1}{2}\Big(H(\Delta L_i^W) + H(L_{i,0}^W)\Big) \leq H(L_{i,1}^W)$ 与 $\frac{1}{2}\Big(H(\Delta L_i^W) + H(L_{i,1}^W)\Big) \leq H(L_{i,0}^W)$，进一步可得

$$H(\Delta L_i^W) \leq \frac{1}{2}\Big(H(L_{i,0}^W) + H(L_{i,1}^W)\Big) \leq \max\Big(H(L_{i,0}^W), H(L_{i,1}^W)\Big) \leq H(L_{i,0}^W, L_{i,1}^W). \tag{3-5}$$

定理 3.3 与定理 3.4 相较于之前 Negrea 等[32-33,197]的工作所提出的数据依赖泛化上界的最显著改进在于使用 $H(R^W)$ 与 $H(\widetilde{R}_\Delta^W)$ 替代了 $I(W;\mathbf{Z})$、$I(W;U|\widetilde{\mathbf{Z}}_l)$ 与 $I(W;\widetilde{U}|\widetilde{\mathbf{Z}}_s)$。为进一步说明，以留一法设定下的泛化上界为例，通过给定 $\widetilde{\mathbf{Z}}_l$ 时的条件 Markov 链：$U \to \mathbf{Z}_l \to W \to R^W$，并进一步结合数据处理不等式，可证明 $I(R^W;U|\widetilde{\mathbf{Z}}_l)$ 严格小于 $I(W;U|\widetilde{\mathbf{Z}}_l)$ 与 $I(W;\mathbf{Z}_l|\widetilde{\mathbf{Z}}_l)$。通过 U 与 $\widetilde{\mathbf{Z}}_l$ 之间的独立性，可得 $I(R^W;U) \leq I(R^W;U) + I(U;\widetilde{\mathbf{Z}}_l|R^W) = I(R^W;U|\widetilde{\mathbf{Z}}_l) + I(U;\widetilde{\mathbf{Z}}_l) = I(R^W;U|\widetilde{\mathbf{Z}}_l)$。类似地，通过利用 $\widetilde{\mathbf{Z}}_l$ 与 W 在给定 $\mathbf{Z}_l$ 时的条件独立性，可证明 $I(W;\mathbf{Z}_l|\widetilde{\mathbf{Z}}_l) \leq I(W;\mathbf{Z}_l|\widetilde{\mathbf{Z}}_l) + I(W;\widetilde{\mathbf{Z}}_l) = I(W;\widetilde{\mathbf{Z}}_l|\mathbf{Z}_l) + I(W;\mathbf{Z}_l) = I(W;\mathbf{Z}_l)$。当使用确定型学习算法 $\mathcal{A}: \mathbf{Z}_l \mapsto W$（例如，使用全梯度下降或固定种子的随机梯度下降）时，R^W 的随机性主要来源于 U，因而有 $H(R^W|U) \approx 0$。结合上述结果，有

$$H(R^W) \approx I(R^W;U) \leq I(R^W;U|\widetilde{\mathbf{Z}}_l) \leq I(W;U|\widetilde{\mathbf{Z}}_l) \leq I(W;\mathbf{Z}_l|\widetilde{\mathbf{Z}}_l) \leq I(W;\mathbf{Z}_l). \tag{3-6}$$

上述观察结果验证了本小节的观点，即通过引入损失熵 $H(R^W)$ 替代 $I(W;\mathbf{Z})$、$I(W;U|\widetilde{\mathbf{Z}}_l)$ 与 $I(W;\widetilde{U}|\widetilde{\mathbf{Z}}_s)$，上述泛化上界相较于现有结果在紧致性方面有了显著改进。类似结论同样适用于 $H(\widetilde{R}_\Delta^W)$。此外，这些现有理论结果在应用于现代深度学习模型时将面对与之前数据无关上界中相似的可计算性挑战，且相较于计算 $I(X;T^w|Y)$ 而言更为严重，因为对于现代深度神经网络而言，W 与 $\mathbf{Z}_l$ 的维度通常远高于 X 或 T^w。相比之下，本小节的泛化上界仅涉及一维随机变量的熵度量，从而可在实际应用中高效计算。

定理 3.3 与定理 3.4 的适用范围十分广泛，可用于模型 W 训练过程涉及数据集 $\mathbf{Z}$ 的相关情形。其相较于数据无关情形下的泛化理论结果（定理 3.1 与定理 3.2）显著扩大了应用范围，包括有监督学习、无监督学习、迁移学习等等常见学习场景。对于定理 3.3 与定理 3.4 而言，其关键信息度量 $H(R^W)$ 或 $H(\widetilde{R}^W)$ 均可进一步分解为 $H(R^W) \leq H(R_{\mathbf{Z}}^W) + H(R_{\overline{\mathbf{Z}}}^W)$，其中 $R_{\mathbf{Z}}^W$ 与 $R_{\overline{\mathbf{Z}}}^W$ 分别代表训练样本损失与测试样本损失的集合。直观而言，学习算法的目标在于最小化训练损失，从而最小化 $H(R_{\mathbf{Z}}^W)$。对于拟合良好的深度学

习模型，所有训练样本损失都应趋于 0。同时，$H(R_{\mathbf{Z}}^{W})$ 指示了模型过拟合的程度，对应于模型对测试样本给出错误答案的情况。这指出了一种在训练与测试损失熵之间存在的权衡以实现最佳的泛化性能。

除上述分析中呈现的主要定理之外，其所采用的证明技术也同样值得探讨。具体而言，观察到对于给定 U 时的条件 Markov 链 $W \to R \to \overline{\mathrm{gen}}$，其中 R（即损失联合熵，如 R^{W} 或 $\widetilde{R}_{\Delta}^{W}$）可作为信息从假设 W 到验证误差 $\overline{\mathrm{gen}}$ 的传递“瓶颈”。通过建立随机变量 R 的典型子集空间 $\mathcal{R}_{\varepsilon}$，可保证同时满足 $\mathbb{P}(R \notin \mathcal{R}_{\varepsilon}) < \delta$ 与 $|\mathcal{R}_{\varepsilon}| = O(e^{H(R)})$。进一步系统性地枚举该典型子集内的每个元素 $r \in \mathcal{R}_{\varepsilon}$，可以有效地将 $\overline{\mathrm{gen}}$ 与 W 解耦，并基于仅利用 U 的随机性的集中不等式构建相关泛化上界。进一步地，通过对 $\mathcal{R}$ 中的每一个元素 r 取联合上界，可建立样本复杂度分析与信息论分析之间的联系，最终得出上述泛化理论结果。虽然本小节主要针对留一法与超样本设定下的泛化问题，但此种技术同样可进行扩展以适应其他应用场景，例如在从总共包含 m 个样本的超样本数据集中随机选择 $n < m$ 个训练样本的学习场景。

3.3.3 快速收敛率的泛化上界

本小节将深入探讨加权验证误差 $\overline{\mathrm{gen}}_{C}(W, \widetilde{\mathbf{Z}}_{s}, \widetilde{U}) = L_{\overline{\mathbf{Z}}_{s}}(W) - (1+C)L_{\mathbf{Z}_{s}}(W)$，其中 C 为一个选定的正常数。通过引入此类加权误差，能够建立具备快速收敛率的泛化误差上界，其收敛率可达到 $1/n$，而非传统上界中的 $1/\sqrt{n}$[31,101]。目前的理论结果通常对所有训练损失均采用统一的 C 值[51,130]，而在实际实验过程中可观察到，对于拟合良好的深度学习模型，单个训练样本的损失值往往表现出长尾分布：大多数训练样本的损失聚集在接近 0 的位置，而少数样本在训练过程结束后仍然表现出相对较高的损失，整体的经验风险则受到这些少数样本的显著影响。受此经验的启发，以下为每个训练样本对应的损失 $L_{i,\widetilde{U}_i}^{W}$ 设置了不同的 C_i 值，从而推导出严格更紧的泛化上界。

定理 3.5 给定任意 $\kappa \geq 0$、$\lambda, \gamma \in (0,1)$ 与 $\delta > 0$，若 $\kappa \geq B^{W,\widetilde{\mathbf{Z}}_{s}}$，则以置信度 $1-\delta$，有

$$\overline{\mathrm{gen}}\left(W, \widetilde{\mathbf{Z}}_{s}, \widetilde{U}\right) \leq \frac{1}{n}\sum_{i=1}^{n} C_i L_{i,\widetilde{U}_i}^{W} + G_1^{W}\frac{H_{1-\lambda}(\widetilde{R}^{W}) + G_2^{W}}{n}, \quad \begin{cases} G_1^{W} = \frac{1}{\eta} = \frac{2\kappa}{\gamma \log 2} \\ G_2^{W} = \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \log\left(\frac{4}{\delta}\right) \\ C_i = -\frac{\log\left(2 - e^{2\eta \widehat{L}_i^{W}}\right)}{2\eta \widehat{L}_i^{W}} - 1 \end{cases} \tag{3-7}$$

其中对于任意 $i \in [1,n]$，定义 $\widehat{L}_i^{W} = \max(L_{i,0}^{W}, L_{i,1}^{W})$。

在训练损失接近 0 的插值模式下，加权验证误差可退化为其原始的非加权形式。因此，当经验风险趋于 0 时，通过取 $\gamma \to 1$，上述定理可达到 $1/n$ 的收敛速率。这一特性使得具有快速收敛率的相关泛化上界在经验风险较小或等于 0 时尤为高效。反之，当

存在较大的训练损失时，权重 C_i 可根据损失值 $\widehat{L}_i^W$ 进行自适应调整，从而赋予了上述上界极高的灵活性，可以适应多种不同损失分布。相比之下，若对于所有训练损失均应用统一的常数 C，则其必须满足 $C \geq \sup_{i\in[1,n]} C_i$，从而在面对非差值情形时将导向严格更松的上界估计结果。

此外，注意到联合熵 $H(\widetilde{R}^W)$ 包含了所有的训练与测试样本。利用信息熵的次可加性，该联合熵可进一步分解为所有单个训练或测试样本的损失熵之和，从而使该上界能够在实际应用中直接量化计算。相比之下，先前工作中具备快速收敛率的高概率泛化上界[51]采用了互信息 $I(\widetilde{R}^W;\widetilde{U}|\widetilde{\mathbf{Z}}_s)$ 作为关键泛化度量。由于 $\widetilde{R}^W$、$\widetilde{U}$ 与 $\widetilde{\mathbf{Z}}_s$ 均为高维随机变量，该上界在实践中不具备可计算性。

然而，上述定理 3.5 依赖于数据集中的最大样本损失 $B^{W,\widetilde{\mathbf{Z}}_s}$，随之呈线性增长关系。在涉及长尾损失分布（如交叉熵）的学习场景中，这一最大样本损失系数将远大于定理 3.3 中的次高斯系数或定理 3.4 中的 L_2 范数。因此，上述上界相较于之前结果由改进收敛率带来的估计紧致度提升在实际中将被显著削弱。

基于这一观察结果，本小节进一步探讨了 $\kappa < B^{W,\widetilde{\mathbf{Z}}_s}$ 下的泛化上界改进。这启发了阈值策略以收紧上述的快速收敛率泛化上界。对于任意阈值 $\kappa > 0$ 与样本损失值 L，容易验证 $L = L^{\kappa} + L^{-\kappa}$，其中 $L^{\kappa} = \min(L,\kappa)$，$L^{-\kappa} = \max(L-\kappa, 0)$。综合定理 3.4 与定理 3.5 的证明策略，可推导出同时具备快速收敛率与适当 L_2 缩放因子的泛化上界：

定理 3.6 给定任意 $\kappa > 0$、$\gamma, \lambda_1, \lambda_2 \in (0,1)$ 与 $\delta > 0$，则以置信度 $1-\delta$，有

$$\overline{\mathrm{gen}}\Big(W, \widetilde{\mathbf{Z}}_s, \widetilde{U}\Big) \leq \frac{1}{n}\sum_{i=1}^{n} C_i L_{i,\widetilde{U}_i}^{W,\kappa} + \widetilde{G}_1^W \frac{H_{1-\lambda_1}(\widetilde{R}^{W,\kappa}) + \widetilde{G}_2^W}{n} + \widetilde{G}_3^W \sqrt{\frac{H_{1-\lambda_2}(\widetilde{R}_{\Delta}^{W,-\kappa}) + \widetilde{G}_4^W}{n}},$$

其中 $\widetilde{R}^{W,\kappa} = \{L_{i,0}^{W,\kappa}, L_{i,1}^{W,\kappa}\}_{i=1}^n$，$\widetilde{R}_{\Delta}^{W,-\kappa} = \{\Delta L_i^{W,-\kappa}\}_{i=1}^n$，$\Delta L_i^{W,-\kappa} = L_{i,1}^{W,-\kappa} - L_{i,0}^{W,-\kappa}$ 且

$$\widetilde{G}_1^W = \frac{1}{\eta} = \frac{2\kappa}{\gamma \log 2}, \quad \widetilde{G}_2^W = \frac{1}{\lambda_1}\log\left(\frac{2}{\delta}\right) + \log\left(\frac{8}{\delta}\right), \quad C_i = -\frac{\log\left(2 - e^{2\eta \widehat{L}_i^{W,\kappa}}\right)}{2\eta \widehat{L}_i^{W,\kappa}} - 1,$$
$$\widetilde{G}_3^W = \sqrt{\frac{2}{n}\sum_{i=1}^{n}\left(\Delta L_i^{W,-\kappa}\right)^2}, \quad \widetilde{G}_4^W = \frac{1}{\lambda_2}\log\left(\frac{2}{\delta}\right) + \log\left(\frac{4}{\delta}\right).$$

定理 3.6 通过引入自定义阈值 κ，将每个样本损失划分为 $L = L^{\kappa} + L^{-\kappa}$。通过结合定理 3.4 与定理 3.5 中的证明路线，可为以上两项分别推导泛化上界。由此，首个损失组成部分 L^{κ} 中初始的最大损失值系数 $B^{W,\widetilde{\mathbf{Z}}_s}$ 被替换为自定义阈值 κ，而第二个组成部分 $L^{-\kappa}$ 的对应系数则替换为具有更慢增长速度的损失差异 L_2 范数。后续实验结果表明，该泛化上界相较于定理 3.4 与定理 3.5 均有所改进。值得一提的是，这种基于阈值策略的证明技术也可应用于 Wang 等[130]的分析结果，能够在面对长尾损失分布时得到更紧的期望泛化误差上界。此外，训练损失的联合熵 $H(\widetilde{R}_{\mathbf{Z}}^W)$ 也可作为一种更紧的替代泛化

度量，用以替代 Wang 等[130]工作中用于推导快速收敛率期望泛化上界的损失方差或损失平坦度相关度量。

3.3.4 连续型损失函数的离散化

在实践中，人们通常通过连续型损失函数优化神经网络模型（例如交叉熵）。虽然这些损失值在计算机系统中通过浮点数机制储存，因而可视为实际上的离散变量，但在实践中无法通过机器精度的分箱大小估计相关变量的熵，因此以上方法在实际中不可行。本小节提出了一种针对连续型损失变量的离散化方法，它适用于任意分箱大小，并可结合上述泛化分析得到有效的面向连续型损失函数的泛化上界。设 $b>0$ 为分箱大小，$\varphi_b(x)$ 是以 b 为基的舍入函数：

$$\varphi_b(x)=b\times\arg\min_{i\in\mathbb{N}}|ib-x|. \tag{3-8}$$

对于训练损失，其由离散化引入的近似误差是可直接计算的。因此，这里仅考虑测试风险上的离散化误差：

定理 3.7 给定任意 $w\in\mathcal{W}$ 与 $b>0$，设 $\{Z_i\}_{i=1}^n\sim\mu^n$ 为独立同分布数据样本，$L_i=\ell(w,Z_i)$ 为对应损失值，$\{D_i\}_{i=1}^n\sim\mathrm{Unif}([-\frac{b}{2},\frac{b}{2}]^n)$ 为独立同分布均匀变量。则以置信度 $1-\delta$，有

$$\frac{1}{n}\sum_{i=1}^n L_i-\frac{1}{n}\sum_{i=1}^n\varphi_b(L_i+D_i)\le b\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}. \tag{3-9}$$

证明： 设 $\widehat{L}_i=L_i-\varphi_b(L_i+D_i)$，则容易验证 $\mathbb{E}_{D_i}[\widehat{L}_i]=0$ 且 $\widehat{L}_i\in[-b,b]$，即 $\widehat{L}_i$ 满足 b-次高斯性。由于 $\{L_i\}_{i=1}^n$ 与 $\{D_i\}_{i=1}^n$ 均为独立同分布变量，因此可得 $\frac{1}{n}\sum_{i=1}^n\widehat{L}_i$ 满足 $\frac{b}{\sqrt{n}}$-次高斯性。故有

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n\widehat{L}_i-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\widehat{L}_i\right]\ge\varepsilon\right)\le\exp\left(-\frac{n\varepsilon^2}{2b^2}\right).$$

通过令 δ 等于上述不等式右侧，则以置信度 $1-\delta$，有

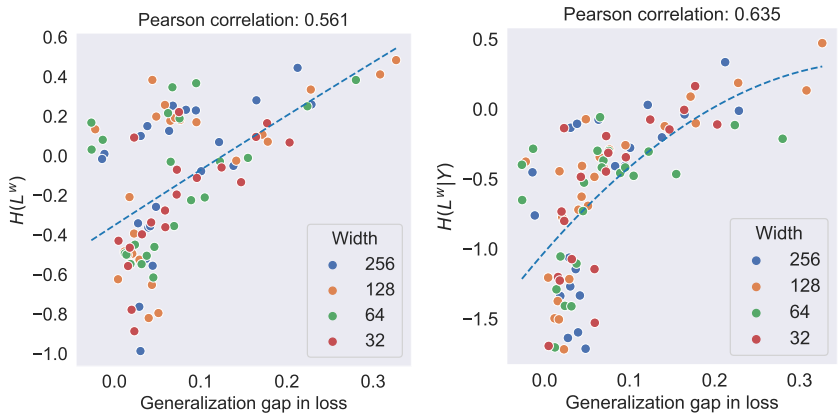
$$\frac{1}{n}\sum_{i=1}^n\widehat{L}_i\le b\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}. \qquad \blacksquare$$

因此，通过为每项测试损失添加随机扰动 $D_i\sim\mathrm{Unif}([-\frac{b}{2},\frac{b}{2}])$ 并四舍五入至最近的分箱，即可完成损失值的离散化过程，且对验证误差不会造成显著影响。通过这种策略，便可利用本小节中面向离散化损失的泛化上界得到相关上界估计，并进一步结合上述定理 3.7 获得面向连续型损失函数的泛化上界。需要注意的是，在本节的主要结果（定理 3.3 - 定理 3.6）中，在给定模型 W 与样本数据 $\mathbf{Z}$ 后，其对应损失值并不需要为确定值。因此，在引入外部随机性 D_i 后，此类泛化上界在 W，$\widetilde{\mathbf{Z}}_s$（或 $\widetilde{\mathbf{Z}}_l$），$\widetilde{U}$（或 U）和

D_i 的抽样概率下仍然有效。

3.4 实验分析

本节通过实验验证对比了本章中构建的泛化上界与先前工作[46,51]中的高概率泛化上界。为此，本节设计了两组实验：首先，使用简单 MLP 网络作为分类器，在模拟二维高斯数据集上评估数据无关的泛化上界，其实验设置与 Kawaguchi 等[46]的工作相同。其次，在真实的图像分类数据集（在 MNIST[200]上训练 4 层 CNN，在 CIFAR10[201]上训练 ResNet-50[202]）上针对现代复杂神经网络评估数据依赖的泛化上界，相关实验设置遵循 Harutyunyan 等[51,58,130]的工作。以下实验将使用交叉熵损失 $\mathcal{L}_{\mathrm{CE}}$ 量化泛化误差，并通过经验风险最小化准则优化假设 W。



(a) $H(L^w)$ (b) $H(L^w|Y)$

图 3-1 泛化误差与损失熵间的 Pearson 相关性分析

3.4.1 模拟数据上的泛化指标对比

对于首个分类任务，本小节使用了模拟二维高斯数据集并通过变分编码器与确定型分类器构建神经网络。遵循 Kawaguchi 等[46]的实验设定，共在不同设置下训练了 216 个不同模型，其中涵盖了不同的模型架构、权重衰减率、数据集抽样和随机种子。利用重参数化技巧[203]，该网络可以端到端的方式进行优化。其中，$I(X;T^w)$ 与 $I(X;T^w|Y)$ 的值可通过 Monte-Carlo 采样估计子近似计算，$I(W;\mathbf{Z})$ 的值则可通过由 SWAG 方法建模的后验分布计算[204-205]。本章提出的条件 $H(L^w|Y)$ 或非条件 $H(L^w)$ 损失熵可通过简单的核密度估计方法直接近似。此处使用了高斯核函数，并根据经验法则确定其核宽度。

为了实证评估不同度量对泛化性能的预测能力，本小节遵循先前工作[46,206]的方法并采用多种相关性分析技术，包括 Pearson、Spearman 与 Kendall 相关性系数。如图 3-1 所示，误差熵 $H(L^w)$ 和 $H(L^w|Y)$ 与真实泛化误差之间存在强相关性。如表 3-1 所示，误

表 3-1 不同度量与泛化误差间的相关性分析

度量	Spearman	Pearson	Kendall
参数数量	-0.0576	-0.0294	-0.0402
$\|W\|_F$	-0.2172	-0.0871	-0.1374
$\widehat{I}(X;T^w)$	0.1816	0.2878	0.1280
$\widehat{I}(X;T^w\|Y)$	0.1749	0.3167	0.1129
$\bar{I}(X;T^w)$	0.1648	0.3712	0.1223
$\bar{I}(X;T^w\|Y)$	0.2293	0.3842	0.1515
$\bar{I}(\mathbf{Z};W)$	0.0020	0.0211	0.0074
$\bar{I}(\mathbf{Z};W)+\widehat{I}(X;T^w)$	0.0178	0.0211	0.0178
$\bar{I}(\mathbf{Z};W)+\widehat{I}(X;T^w\|Y)$	0.0163	0.0211	0.0167
$\bar{I}(\mathbf{Z};W)+\bar{I}(X;T^w)$	0.0135	0.0212	0.0162
$\bar{I}(\mathbf{Z};W)+\bar{I}(X;T^w\|Y)$	0.0164	0.0211	0.0167
$\widetilde{I}(\mathbf{Z};W)+\widehat{I}(X;T^w)$	0.1104	0.1401	0.0794
$\widetilde{I}(\mathbf{Z};W)+\widehat{I}(X;T^w\|Y)$	0.2253	0.3177	0.1567
$\widetilde{I}(\mathbf{Z};W)+\bar{I}(X;T^w)$	0.2684	0.3928	0.1912
$\widetilde{I}(\mathbf{Z};W)+\bar{I}(X;T^w\|Y)$	0.3015	0.4130	0.2085
$H(L^w)$	<u>0.5767</u>	<u>0.5611</u>	<u>0.4037</u>
$H(L^w\|Y)$	**0.7088**	**0.6350**	**0.5251**

差熵相比于其他度量（包括参数数量、假设的 F-范数、信息瓶颈 $I(X;T^w)$ 度量与输入—输出互信息 $I(W;\mathbf{Z})$）具备更强的相关性。在表 3-2 中，进一步对比了 20% 标签噪声下的相关性分析结果，可见损失熵度量始终是泛化性能的良好指标。此外，$H(L^w|Y)$ 与 $H(L^w)$ 的对比结果表明，当标签空间基数 $|\mathcal{Y}|$ 有限时，定理 3.2 将比定理 3.1 更有效。

3.4.2 真实数据上的泛化上界对比

对于真实场景下的泛化性能评估，本小节遵循了 Harutyunyan 等[58]的实验设定。具体操作为在二值化 MNIST 数据集（仅包含数字 4 和 9）上训练 4 层卷积神经网络（CNN），并于预训练 ResNet-50 模型的基础上在 CIFAR10 数据集上进行微调。在每项超样本（留一法）设置下的实验中，抽取 k_1 个 $\widetilde{\mathbf{Z}}_s$（$\widetilde{\mathbf{Z}}_l$）的样本，即从相应的数据集中随机抽取 $2n$（$n+1$）个样本。对于每个 $\widetilde{\mathbf{Z}}_s$（$\widetilde{\mathbf{Z}}_l$），继而进行 k_2 次不同的训练/测试数据划分 $\widetilde{U}$（U），故总共有 $k_1 \times k_2$ 次独立训练过程。通过上述离散化策略，以 1.0 的分箱大小离散化损失值。以下选择 $\delta = 0.5$ 以模拟相关泛化上界的期望值。

在目前工作中，已知的最紧信息论高概率泛化估计结果来自于 Hellström 等[51]的工作。其上界中的关键度量取自于高维随机 KL 散度，鉴于其不可计算性，可取其在期望意义下的 KL 散度度量，其等价于 $I(\widetilde{R}^W;\widetilde{U}|\widetilde{\mathbf{Z}}_s)$。然而，该互信息度量仍包含高维随机变量，故在实际中不可计算。为此，将取其下界估计：

定理 3.8 $\sum_{i=1}^{n} I(L_{i,0}^W, L_{i,1}^W; \widetilde{U}_i) \le I(\widetilde{R}^W;\widetilde{U}|\widetilde{\mathbf{Z}}_s)$.

表 3-2 标签噪声情形下，不同度量与泛化误差间的相关性分析

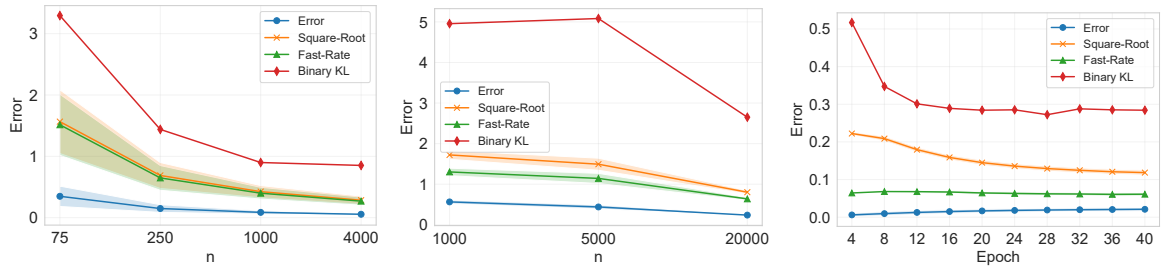
度量	Spearman	Pearson	Kendall
参数数量	0.4944	0.2770	0.3985
$\|W\|_F$	0.4944	0.2680	0.3985
$\widehat{I}(X;T^w)$	0.4799	0.6232	0.2941
$\widehat{I}(X;T^w\|Y)$	0.4923	0.6185	0.3203
$\bar{I}(X;T^w)$	0.1496	0.0190	0.0196
$\bar{I}(X;T^w\|Y)$	0.2198	0.0495	0.0980
$\bar{I}(\mathbf{Z};W)$	0.6065	0.5692	0.4633
$\bar{I}(\mathbf{Z};W)+\widehat{I}(X;T^w)$	0.6140	0.6417	0.4248
$\bar{I}(\mathbf{Z};W)+\widehat{I}(X;T^w\|Y)$	0.6202	0.6406	0.4510
$\bar{I}(\mathbf{Z};W)+\bar{I}(X;T^w)$	0.3808	0.1378	0.2549
$\bar{I}(\mathbf{Z};W)+\bar{I}(X;T^w\|Y)$	0.4138	0.1648	0.2810
$\widetilde{I}(\mathbf{Z};W)+\widehat{I}(X;T^w)$	0.6223	0.5692	0.4510
$\widetilde{I}(\mathbf{Z};W)+\widehat{I}(X;T^w\|Y)$	0.6223	0.5692	0.4510
$\widetilde{I}(\mathbf{Z};W)+\bar{I}(X;T^w)$	0.5666	0.5685	0.3987
$\widetilde{I}(\mathbf{Z};W)+\bar{I}(X;T^w\|Y)$	0.5810	0.5707	0.4118
$H(L^w)$	<u>0.8782</u>	<u>0.8679</u>	<u>0.7647</u>
$H(L^w\|Y)$	**0.9030**	**0.8915**	**0.7778**

证明： 简便起见，以下使用 $\widetilde{U}_{1:n}$ 作为 $\{\widetilde{U}_i\}_{i=1}^n$ 的简写：

$$
\begin{aligned}
I(\widetilde{R}^W;\widetilde{U}|\widetilde{\mathbf{Z}}_s) &= I(\widetilde{R}^W;\widetilde{U}|\widetilde{\mathbf{Z}}_s)+I(\widetilde{U};\widetilde{\mathbf{Z}}_s)=I(\widetilde{R}^W;\widetilde{U})+I(\widetilde{U};\widetilde{\mathbf{Z}}_s|\widetilde{R}^W)\\
&\geq I(\widetilde{R}^W;\widetilde{U})=I(\widetilde{R}^W;\widetilde{U}_1)+I(\widetilde{R}^W;\widetilde{U}_{2:n}|\widetilde{U}_1)\\
&=I(\widetilde{R}^W;\widetilde{U}_1)+I(\widetilde{R}^W;\widetilde{U}_{2:n})-I(\widetilde{U}_{2:n};\widetilde{U}_1)+I(\widetilde{U}_{2:n};\widetilde{U}_1|\widetilde{R}^W)\\
&=I(\widetilde{R}^W;\widetilde{U}_1)+I(\widetilde{R}^W;\widetilde{U}_{2:n})+I(\widetilde{U}_{2:n};\widetilde{U}_1|\widetilde{R}^W)\\
&\geq I(\widetilde{R}^W;\widetilde{U}_1)+I(\widetilde{R}^W;\widetilde{U}_{2:n})\geq\cdots\geq\sum_{i=1}^n I(\widetilde{R}^W;\widetilde{U}_i)\\
&=\sum_{i=1}^n I(L_{i,0}^W,L_{i,1}^W;\widetilde{U}_i)+I(\widetilde{R}^W\setminus\{L_{i,0}^W,L_{i,1}^W\};\widetilde{U}_i|L_{i,0}^W,L_{i,1}^W)\geq\sum_{i=1}^n I(L_{i,0}^W,L_{i,1}^W;\widetilde{U}_i).\ \blacksquare
\end{aligned}
$$

上述上界通常称为二值 KL 上界（Binary KL）。以下将其与定理 3.4 中的平方根泛化上界（Square-Root）与定理 3.6 中的快速率泛化上界（Fast-Rate）进行对比。相关泛化上界中的超参数将通过优化算法选择。具体而言，本小节使用 L-BFGS-B 算法选择最优的 λ，使用 Nelder-Mead 算法选择最优的 γ 与 κ，并使用 brentq 算法求解 Hellström 等[51]上界中的二值 KL 散度。值得注意的是，上述超参数不应直接针对泛化上界右侧进行优化，因为它们被假定为常数，其取值应独立于 W、$\widetilde{\mathbf{Z}}_s$ 或 $\widetilde{U}$ 的选择。在实验中，可通过优化上述上界在多次实验上的期望值以获得理想的超参数选择。

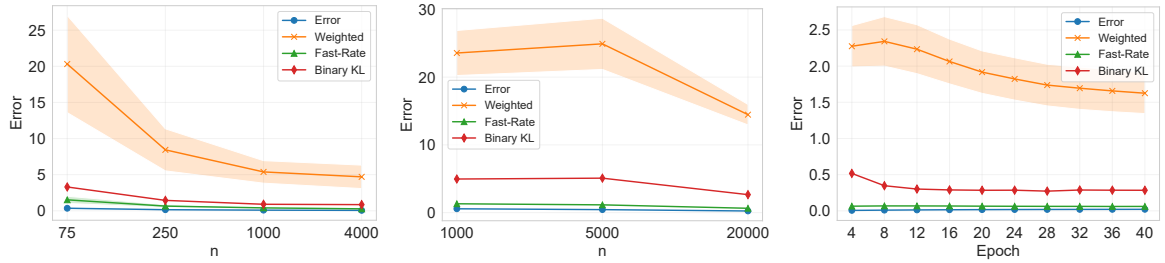
最终对比结果如图 3-2 所示。可见，这些上界均可有效预测泛化误差的变化趋势。



(a) MNIST, Adam (b) CIFAR10, SGD (c) MNIST, SGLD

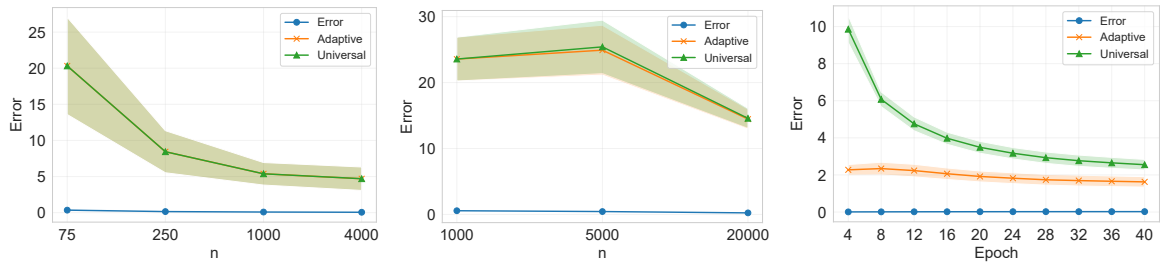
图 3-2 不同学习场景下的超样本泛化误差上界对比

在三种不同的学习场景中，本章中的平方根上界与快速率上界均显著优于二值 KL 上界的下界近似。这一观察结果验证了之前的分析，即本章中的上界能够自然适应长尾损失分布，而二值 KL 上界则受到最大损失值的显著影响。此外，快速收敛率的泛化上界在更复杂的学习场景中（如 CIFAR10 和 SGLD）始终优于平方根上界。这为上述阈值策略的有效性提供了实证验证。



(a) MNIST, Adam (b) CIFAR10, SGD (c) MNIST, SGLD

图 3-3 加权泛化上界（定理 3.5）与其他上界的对比

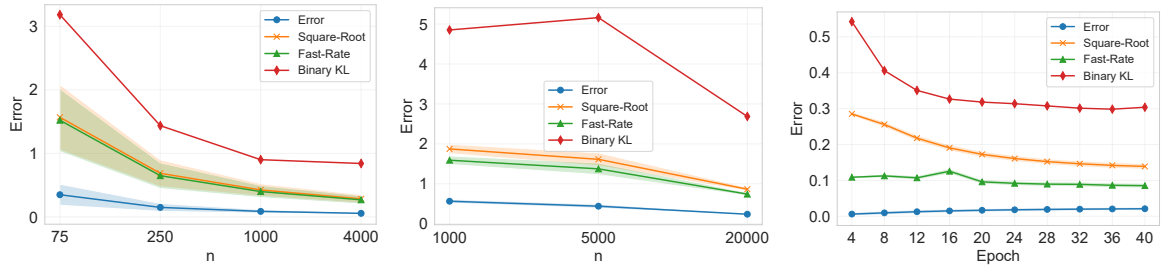


(a) MNIST, Adam (b) CIFAR10, SGD (c) MNIST, SGLD

图 3-4 自适应 C_i 与统一 C_i 值的上界对比

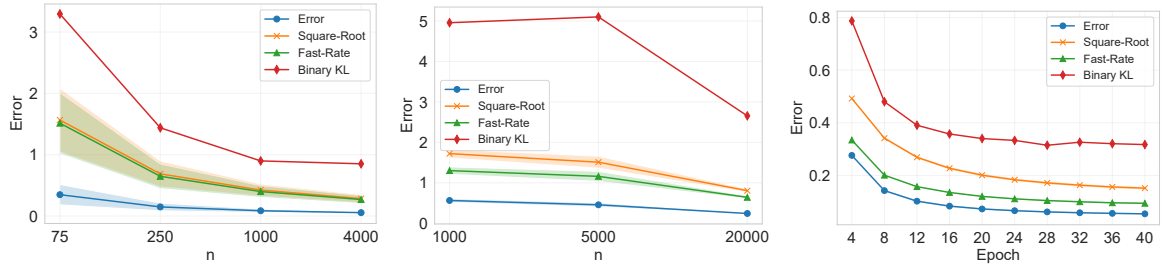
随后，将评估本章中所提出的阈值方法的有效性。如图 3-3 所示，上述快速率泛化上界（定理 3.6）相较于加权泛化上界（定理 3.5）显著改进了紧致性。若不采用阈值方法，则快速率泛化上界无法取得比二值 KL 上界更优的泛化估计结果。类似地，图 3-4 展示了为每个单独训练损失自适应选择系数 C_i 的重要性。可见，这种策略在训练早期

尤为有效，显著改进了训练损失接近（但不等于）设定阈值时的泛化上界。对于拟合良好的网络，当训练损失趋近于零时，自适应选择系数 C_i 所带来的改进将不再显著。



(a) MNIST, Adam　　(b) CIFAR10, SGD　　(c) MNIST, SGLD

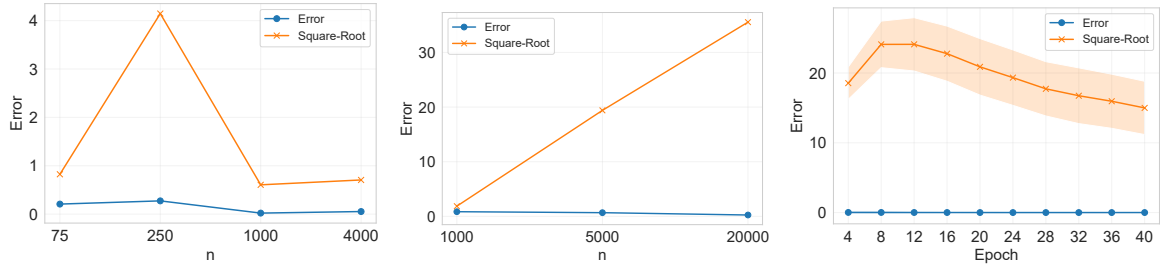
图 3-5 0.6 分箱大小下的泛化误差上界对比



(a) MNIST, Adam　　(b) CIFAR10, SGD　　(c) MNIST, SGLD

图 3-6 不同学习场景下的超样本测试风险上界对比

为评估分箱大小对相关可视化结果的影响，图 3-5 中展示了分箱大小为 0.6 时的泛化上界对比。一般而言，减小分箱大小将导向更大的损失熵度量，但同时导向较低的离散化误差。因此，需具体考虑两者间的权衡以获得最紧的泛化上界。此外，图 3-6 中可视化了不同学习场景下测试风险的上界对比。



(a) MNIST, Adam　　(b) CIFAR10, SGD　　(c) MNIST, SGLD

图 3-7 不同学习场景下的留一法泛化误差上界对比

最后，图 3-7 中展示了留一法设定下的泛化上界（定理 3.3）与真实泛化误差的对比。虽然此上界对于较大的 n 值无法很好地预测泛化误差的变化，但值得一提的是，这是在留一法设定下的首个可计算的高概率泛化上界。

3.5 本章小结

本章基于新型低维信息论泛化度量：损失熵，引入了一系列高概率的信息论泛化理论结果。损失熵从根本上解决了目前信息论泛化上界的不可计算问题，其仅包含一维随机变量，故可通过核密度估计、分箱方法等直接近似计算。在此之上，成功将其拓展至 Kawaguchi 等[46]工作中的数据无关泛化理论，显著提升了相关泛化上界的紧致性与可计算性，并可为最小化误差熵准则提供全新的理论见解。进一步地，对于数据依赖泛化上界，在留一法与超样本泛化分析框架下改进了现有依赖于高维信息度量的泛化理论结果，构建了基于损失熵的高概率泛化误差上界估计，使面向随机学习算法的泛化误差精确数值刻画成为可能。最后，在多项经典深度学习任务上验证了相关理论结果，发现了损失熵与泛化性能之间的强相关性，并验证了本章中的泛化上界相较于现有理论[51]更能够提供准确的泛化误差估计结果。

本章的研究工作发表于机器学习顶级会议、清华推荐 A 类学术会议 International Conference on Learning Representations，论文题目为“Rethinking Information-theoretic Generalization: Loss Entropy Induced PAC Bounds”。

4 面向多点损失的一致信息论泛化分析框架

4.1 引言

随着深度学习技术的发展，面向传统有监督学习的单点学习框架无法涵盖以对比学习[207-209]为代表的多点学习范式。此类学习场景通常对应多点损失函数，即每项损失均需要多个数据样本进行评估，而非传统单点学习中损失与样本一一对应。因此，目前面向单点损失函数构建的泛化分析理论不再适用于此类多点学习场景。其他常见的多点学习应用还包括深度度量学习[210-212]、AUC 最大化[213-214]和排序算法[215-216]等等。此类学习场景不但在实际任务中得到了广泛应用，其相关理论基础也在经典一致收敛性[217-219]与算法稳定性[179,183,220]视角下得到了一定探索。然而，现有的多点学习泛化研究工作主要局限于双点[221-223]与三点[224]学习场景中，对四点[225]或更高阶[207,211]情形的探索仍处于起步阶段。此外，这些分析方法常常依赖于假设空间复杂性或强假设（如 Lipschitz 连续性、平滑性与凸性），在面对深度神经网络时往往导向无意义或不可计算的泛化上界。

近期，基于信息论的泛化分析工作[28]为分析随机迭代学习算法的泛化性能提供了切实可行的替代方案[33-34,55,226]。基于信息论的泛化分析技术可同时将数据样本与假设的概率分布纳入考虑，因此能够结合具体学习算法分析其独有性质，从而在面对现代深度学习模型时具备更强的理论可解释性。同时，其前提假设相较于算法稳定性等分析方法更为宽松，从而具有更为广阔的应用前景。信息论泛化研究的最新进展通过利用网络预测值[58]、样本损失[51]或损失差异[130]等相关信息度量，在超样本框架下提供了更精确的泛化误差估计结果[32]。此类上界不仅由于其关键信息度量的低维性质而得以在实际应用中直接量化计算，更可适用于任意规模的深度神经网络，为广泛机器学习模型的泛化性能提供了准确的数值估计。然而，这些分析目前仍局限于单点学习场景，纵使是相对简单的双点学习情形也仍未得到探索。

将此类泛化理论结果拓展至多点学习场景将面临多项挑战。首先，其损失值不再通过独立同分布样本评估，而是通过成对的样本子集进行。这将使不同的损失变量包含重复的训练样本数据，从而失去独立性。因此，其经验风险不再是独立同分布损失变量的平均值，而这一特性对基于输入—输出互信息泛化上界的推导至关重要[28,54]，更是分析随机迭代学习算法泛化性能的基石[33,141]。其次，在超样本框架下推广基于网络预测值的泛化上界也将面临维度爆炸问题，其关键信息度量的维度将随损失函数包含的样本数量 m 呈指数级增长，在面对三点学习等高阶场景时将失去低维性质，导致此类上界最为突出的可计算性在高阶场景下不复存在。

本章中的理论工作突破了上述障碍，为各阶多点学习模型提供了统一的信息论泛

化分析框架。具体而言，本章通过一种自底向上的上界归约技巧克服了拓展相关泛化上界时面临的非独立同分布损失挑战：利用期望算子的线性性质，可将整体泛化误差的上界推导拆分为训练集子集的泛化上界推导，对于每项损失分别构建其独立泛化上界，并利用互信息度量的超可加性 (Superadditivity) 进行上界归约，从而得到整体泛化误差上界。进一步地，本章通过一种针对超样本变量的独立性分解技巧，克服了超样本框架下相关信息度量的维度爆炸问题：利用异或算子不影响二值随机变量独立性的独特性质，可将每项损失对应的 m 维超样本变量拆分为 1 维与 $m-1$ 维子变量，并综合损失差异变量构建了仅包含一维随机变量的互信息泛化上界。其维度将不再随损失函数中的样本数量 m 而增长，从而使相关上界在高阶多点学习场景中仍能保持其可计算性，同时相较于现有上界的简单拓展在紧致性上亦有提升。基于上述泛化分析技术，本章构建了面向多点学习的一致信息论分析框架，其能够将不同阶学习场景同时纳入统一的分析体系，涵盖了包括单点、双点、三点及更高阶情形在内的常见学习场景。同时，这也是首个适用于多点学习场景的信息论泛化上界。最后，在多个模拟与真实数据集上验证了上述理论结果，证实了本章提出的理论泛化上界能够准确刻画多种现代深度学习任务下的泛化误差变化曲线。

总体而言，本章的主要贡献包括：（1）提出了首个面向多点学习的信息论泛化上界，其适用于任意有界损失函数，并通过统一的泛化分析框架涵盖了包括单点、双点、三点及更高阶情形在内的学习范式。（2）通过基于互信息超可加性的上界归约技术，克服了拓展相关上界时面临的非独立同分布损失挑战，成功推广了目前基于输入—输出互信息与条件互信息度量的泛化理论结果，并进一步分析了随机迭代学习算法在多点学习场景下的泛化能力。（3）通过针对超样本变量的独立性分解技术，克服了拓展现有超样本泛化上界时面临的维度爆炸问题，得到了仅包含一维随机变量的互信息泛化上界，且其维度不随样本数量 m 而增长，维持了其在可计算性与紧致性上的优势。

4.2 问题设定与相关背景

本节将常见的多点学习场景纳入统一的泛化分析框架之中，这些场景根据其损失函数所包含的样本数量可分类为：

（1） 单点学习（$m=1$）：交叉熵，均方误差；

（2） 双点学习（$m=2$）：对比损失；

（3） 三点学习（$m=3$）：三元损失；

（4） 四点学习（$m=4$）：四元损失；

（5） 更高阶情形（$m\geq 5$）：N-pair 损失，NT-Xent 损失。

设 $\ell:\mathcal{W}\times\mathcal{Z}^m\mapsto\mathbb{R}^+$ 为损失函数。对于给定假设 $w\in\mathcal{W}$，可定义总体风险为 $L(w)\triangleq\mathbb{E}_{Z_{1:m}}[\ell(w,Z_{1:m})]$，其中 $Z_{1:m}\sim\mu^m$ 是独立同分布采样的一组测试样本。期望总体风险可定

义为 $L = \mathbb{E}_W[L(W)]$。设 P_n^m 为从 n 个不同元素中取出 m 个元素的所有排列情形构成的集合，C_n^m 则为对应组合情形构成的集合。给定一组索引序列 $u = \{u_i\}_{i=1}^m \in [1,n]^m$，设 $Z_u = \{Z_{u_i}\}_{i=1}^m$ 表示通过 u 索引的训练样本序列，则经验风险可定义为 $L_{\mathbf{Z}}(w) \triangleq \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\ell(w,Z_u)$。类似地，定义 $L_n = \mathbb{E}_{W,\mathbf{Z}}[L_{\mathbf{Z}}(W)]$ 为期望经验风险。期望意义下的泛化误差定义为 $\overline{\mathrm{gen}} \triangleq L - L_n$，其量化了经验风险与总体风险间的差异。

Steinke 等[32]的工作首次探索了超样本设定下的泛化分析方法。设 $\widetilde{\mathbf{Z}} = \{\widetilde{Z}_i\}_{i=1}^n \in \mathcal{Z}^{n\times 2}$ 为从数据分布 μ 中独立同分布采样的超样本数据集，其中每个元素 $\widetilde{Z}_i = (\widetilde{Z}_i^0, \widetilde{Z}_i^1)$ 由一对数据样本构成。随后，采样一组二值随机变量 $S = \{S_i\}_{i=1}^n \sim \mathrm{Unif}(\{0,1\}^n)$ 用以划分训练与测试样本：$\widetilde{\mathbf{Z}}_S = \{\widetilde{Z}_i^{S_i}\}_{i=1}^n$ 构成训练数据集，$\widetilde{\mathbf{Z}}_{\overline{S}} = \{\widetilde{Z}_i^{\overline{S}_i}\}_{i=1}^n$ 则构成测试数据集。相应地，超样本设定下的经验风险与总体风险分别定义为 $L_n = \mathbb{E}_{W,\widetilde{\mathbf{Z}},S}[L_{\widetilde{\mathbf{Z}}_S}(W)]$ 与 $L = \mathbb{E}_{W,\widetilde{\mathbf{Z}},S}[L_{\widetilde{\mathbf{Z}}_{\overline{S}}}(W)]$。

设 $\mathsf{B}_m = \{0,1\}^m$ 为长度为 m 的所有二值序列所构成的集合。给定任意 $u \in \mathsf{P}_n^m$ 与 $b \in \mathsf{B}_m$，定义 $\widetilde{Z}_u^b$ 表示由 u 与 b 索引的样本子集 $\{\widetilde{Z}_{u_i}^{b_i}\}_{i=1}^m$，且设 $L_u^b = \ell(W,\widetilde{Z}_u^b)$ 为该样本子集对应的损失值。所有 $b \in \mathsf{B}_m$ 组合对应损失值的集合表示为 $L_u = \{L_u^b\}_{b\in\mathsf{B}_m}$。设 $S_u = \{S_{u_i}\}_{i=1}^m \in \mathsf{B}_m$ 为由 u 索引的超样本变量子集，并设 $\Phi_u = \{S_{u_1} \oplus S_{u_i}\}_{i=2}^m \in \mathsf{B}_{m-1}$，其中 $\oplus$ 为异或运算。给定二值常量 $b \in \{0,1\}$，定义 $b \otimes \Phi_u = (b, \{\Phi_{u_i} \oplus b\}_{i=1}^{m-1}) \in \mathsf{B}_m$。简便起见，将 $0 \otimes \Phi_u$ 与 $1 \otimes \Phi_u$ 分别简写为 Φ_u^- 与 Φ_u^+。进一步地，$L_u^{\Phi_u} = (L_u^{\Phi_u^-}, L_u^{\Phi_u^+})$ 表示一对损失值，$\Delta_u^{\Phi_u} = L_u^{\Phi_u^+} - L_u^{\Phi_u^-}$ 则为对应的损失差异。

4.2.1 常见多点学习场景

对比表征学习（以下简称为对比学习）通过不同样本数据间的表征对比监督以增强机器学习模型的预测性能。对比学习的核心在于对比损失函数的刻画，其常基于编码器网络 $f: \mathcal{X} \mapsto \mathcal{T}$ 所提取的特征表示 T_i 间的相似性计算。其基本原则在于最小化相似样本间的距离，同时最大化相异样本间的距离。具体而言，通过定义相似度指标 $d: \mathcal{T}\times\mathcal{T} \mapsto \mathbb{R}^+$，相似样本将在表征嵌入空间中被拉近（即最小化相似度），而差异较大的样本表征嵌入则会被推远（最大化相似度）。常见的相似性度量包括欧氏距离、余弦相似度等。以下为一种常见的对比损失函数，称为最大边际对比损失：

$$\ell_{\mathrm{contr}}(X_i,X_j) = 1_{Y_i=Y_j}\cdot d(T_i,T_j) + 1_{Y_i\neq Y_j}\cdot \max\{\varepsilon - d(T_i,T_j), 0\}, \tag{4-1}$$

其中 ε 为定义不同类样本之间最小距离的边际超参数。模型通过优化上述损失可学习得到稳健的特征表示，其可随后应用到广泛的下游任务中。进一步地，Schroff 等[227]提出了三元对比损失函数，其通过引入正样本与负样本的概念，允许模型同时优化样本间的相似性与相异性，以此增强对比学习模型的性能：

$$\ell_{\text{tri}}(X_i, X_+, X_-) = \max\{d(T_i, T_+) - d(T_i, T_-) + \varepsilon, 0\}. \tag{4-2}$$

上述损失函数强制要求样本标签满足 $Y_i = Y_+$ 和 $Y_i \neq Y_-$，与本节中基于排列数的问题设定并不直接兼容。以下，对此优化目标做轻微调整以解决此问题：

$$\ell_{\text{tri}}(X_i, X_j, X_k) = \begin{cases} \max\{d(T_i, T_j) - d(T_i, T_k) + \varepsilon, 0\}, & Y_i = Y_j, Y_i \neq Y_k, \\ 0, & \text{其他}. \end{cases}$$

对于上述修改后的对比损失函数，其经验与总体风险均可通过对所有排列情况 $u \in \mathsf{P}_n^3$ 计算对应损失值并取其平均值而表示，从而可自然融入到本节中的泛化分析框架中。Chen 等[225]将三元损失进一步拓展，得到了四元对比损失函数：

$$\ell_{\text{quad}}(X_i, X_j, X_k, X_l) = \begin{cases} \max\{d(T_i, T_j) - d(T_k, T_l) + \varepsilon, 0\}, & Y_i = Y_j, Y_i \neq Y_k, Y_i \neq Y_l, Y_k \neq Y_l, \\ 0, & \text{其他}. \end{cases}$$

此外，由 Sohn 等[211]发展的 N-pair 损失函数适用于任意数量的负样本，并对于取最大值运算引入平滑函数，得到以下对比损失函数：

$$\ell_{\text{n-pair}}(X_{1:m}) = \begin{cases} \log\Big(1 + \sum_{i=3}^{m} \exp(d(T_1, T_i) - d(T_1, T_2))\Big), & Y_1 = Y_2, Y_1 \neq Y_i, \forall i \in [3, m], \\ 0, & \text{其他}. \end{cases}$$

对比表征学习中常见的损失函数还包括 NT-Xent 损失[207]和 InfoNCE 损失[228]等。这些损失函数均可应用于任意数量的对比样本。然而，目前针对对比学习的泛化分析结果仍局限于双点或三点学习。相比之下，本节的泛化分析框架可自然拓展至任意大的 m 值，从而为对比学习泛化提供了更全面、更灵活的分析框架。

深度度量学习专注于量化数据样本之间的相似性，其目标为联合训练编码器 $f: \mathcal{X} \mapsto \mathcal{T}$ 与相似性度量函数 $d: \mathcal{T} \times \mathcal{T} \mapsto \mathbb{R}^+$。深度度量学习的基本原则在于使对于任意给定数据样本 X_i, X_j 及其对应标签 Y_i, Y_j，当两者标签相同时可计算得到较大的相似度 $d(T_i, T_j)$，反之则得到较小的相似度。深度度量学习的损失函数刻画与对比表征学习非常相似，而关键区别在于相似性度量 d 的性质。对于对比表征学习，d 通常是预定义的固定函数，而深度度量学习则将 d 视为训练目标之一，通过训练使其能够分辨不同数据样本间的细微差别。鉴于其概念上的相似性，本节的泛化分析框架同样适用于深度度量学习相关场景。

排序算法旨在同时处理不同数据特征，并基于此预测它们之间的最优排列顺序。此类算法在搜索引擎与推荐系统中有着广泛的应用。目前工作中的常见排序算法可依据

其具体实现方式分为以下三种：

（1） 单点排序：此类方法将为每一数据特征向量计算对应得分，并将其作为后续排序的依据以确定其具体排序。单点排序方法将排序问题视为回归或分类任务，并将每个样本视为独立任务。

（2） 双点排序：双点排序方法通过成对样本间的对比结果确定排序。典型的双点排序模型可表示为$f: \mathcal{X} \times \mathcal{X} \mapsto [0,1]$，其同时接受两个样本作为输入，并输出第一个样本排序高于第二个样本的概率。此类方法关注的是样本间的相对顺序，从而将排序问题转化为一种双点二分类任务。

（3） 序列排序：与单点或双点排序不同，序列排序方法将一次性处理全部样本，其输入为整个样本集合，输出则为这些样本间的排序关系。因此，其损失函数中的样本数量 m 取决于样本集合的具体大小。

现有面向多点学习的泛化上界已经能够应用于单点或双点排序算法的泛化分析，但对于序列排序方法则研究较少，尤其是当 m 较大时的情形。本节首次将多点学习的泛化分析拓展到涵盖任意序列长度的排序方法。

4.3 基于假设的泛化上界

4.3.1 基于输入—输出互信息的泛化上界

Xu 等[28]的先驱工作引入了基于输入—输出互信息的泛化上界，即通过输入数据集 $\mathbf{Z}$ 与输出假设 W 间的互信息建立期望泛化误差的上界估计。这一概念在后续研究中得到了进一步拓展与完善，Bu 等[54,58]通过引入随机子集方法证明了以下结果：

引理 4.1 ([58]，定理 2.2) 假设 $m = 1$ 且 $\ell(\cdot,\cdot) \in [0,1]$，则对于任意 $k \in [1,n]$ 有

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{C}_n^k|} \sum_{u \in \mathsf{C}_n^k} \sqrt{\frac{1}{2k} I(W; Z_u)}. \tag{4-3}$$

在上述上界中设 $k = n$ 可得到一种简化结果：注意到 $|\mathsf{C}_n^k| = 1$，从而可得此上界的收敛速率为 $O(\sqrt{1/n})$。这一结果依赖于给定假设 $w \in \mathcal{W}$ 时，训练损失项满足独立同分布条件的前提假设。在此条件下，经验风险 $L_{\mathbf{Z}}(w)$ 可表示为 n 个独立同分布 $\frac{1}{2}$-次高斯变量的平均值，故而满足 $\frac{1}{2\sqrt{n}}$-次高斯性。然而，这种独立性条件仅在单点学习场景中（$m = 1$）成立。对于多点损失函数（$m > 1$），由于不同损失项将包含重复的训练数据样本，其独立性条件将不再成立。Bu 等[54]首次提出利用单点信息稳定性 $I(W; Z_i)$ 构建泛化误差上界。受此启发，以下将这一概念拓展至多点学习场景中的组信息稳定性：

$$\left| \mathbb{E}_{W,Z_u}[\ell(W, Z_u)] - L \right| \leq \sqrt{\frac{1}{2} I(W; Z_u)}, \tag{4-4}$$

其中 $u \in \mathsf{P}_n^m$ 代表了由 m 个数据样本构成的训练集 $\mathbf{Z}$ 子集。利用独立随机变量互信息的超可加性，组稳定性 (Group Stability) 可作为由输入—输出互信息度量的平均稳定性的替代度量。随后，基于此可构建适用于任意多点学习场景的拓展泛化上界：

定理 4.2 假设 $\ell(\cdot,\cdot) \in [0,1]$，则对于任意 $k \in \left[1, \frac{n}{m}\right]$ 有

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \sqrt{\frac{1}{2k} I(W; Z_u)}. \tag{4-5}$$

这一上界与 Xu 等[28]理论结果的基本原则一致：输出假设 W 对输入数据集 $\mathbf{Z}$ 的依赖性越小，则该学习算法的泛化能力越强。在定理 4.2 中，这一原则得以进一步拓展到多样本损失函数，其适用于任意 $m \geq 1$ 的情形，涵盖了包括双点与三元对比学习在内的多种多点学习范式。假设 $n \bmod m = 0$ 并取 $k = \frac{n}{m}$，上述定理实现了 $O(\sqrt{m/n})$ 的收敛速率，这与引理 4.1 中 $m = 1$ 时的单点学习上界相吻合。该收敛速率同样得到了之前基于一致稳定性的双点[179,220]或三点[224]理论分析结果的验证。此外，上述定理表明收敛速率收到 $\sqrt{m}$ 因子的负面影响，这种影响来源于不同损失变量间的相关性。特别地，该定理首次建立了多点泛化误差与损失函数样本数量 m 之间的联系。

基于 Hellström 等[51]的工作，本小节进一步构建了基于期望经验风险与期望总体风险间二值 KL 散度的泛化误差上界：

定理 4.3 假设 $\ell(\cdot,\cdot) \in [0,1]$，则对于任意 $k \in \left[1, \frac{n}{m}\right]$ 有

$$d(L_n \,\|\, L) \leq \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} I(W; Z_u). \tag{4-6}$$

进一步地，在插值模式下（即 $L_n = 0$ 时）有

$$L \leq \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} I(W; Z_u). \tag{4-7}$$

上述定理表明，当训练风险接近或等于 0 时，该上界能够达到 $O(m/n)$ 的快速收敛率，其相较于此前的 $O(\sqrt{m/n})$ 收敛率有显著改进。移除上界中的平方根有利于在输入—输出互信息度量较小时得到更紧地泛化上界。具体而言，给定任意 $k \in [1, \frac{n}{m}]$，若对于任意 $u \in \mathsf{P}_n^{km}$ 均有 $I(W; Z_u) \leq \frac{k}{2}$，则上述插值模式下的泛化上界相较于定理 4.2 中的平方根上界更紧。这一条件在实际训练场景中通常能够得到满足：在训练集足够大时（即 $n \to \infty$）泛化误差将趋于 0，从而意味着 $I(W; \mathbf{Z}) = o(n)$。

在以上上界中，如何选择合适的 k 值以得到最紧的上界估计是一个值得探究的问题。对于较小的 k 值，上述互信息度量与收敛阶的分母将同时减小。特别地，Harutyunyan 等[58]的研究结果表明引理 4.1 中的泛化上界关于 k 严格非递减，这表明选择最小的 k 值

（即 $k=1$）能够导向最紧的泛化上界。以下将这一结论推广至多点学习情形：

定理 4.4 设 $\varphi:\mathbb{R}\mapsto\mathbb{R}$ 为任意非递减凹函数，则对于任意 $k\in[1,\frac{n}{m}-1]$，有

$$\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\varphi\left(\frac{1}{2k}I(W;Z_u)\right)\leq\frac{1}{\left|\mathsf{C}_n^{km+m}\right|}\sum_{u\in\mathsf{C}_n^{km+m}}\varphi\left(\frac{1}{2(k+1)}I(W;Z_u)\right). \tag{4-8}$$

通过选取 $\varphi(x)=\sqrt{x}$，可以验证 $k=1$ 是最小化定理 4.2 中泛化上界的最优选择。同理，选取 $\varphi(x)=x$ 即可得关于定理 4.3 的相同结论。虽然选择 $k=\frac{n}{m}$ 不利于得到更紧的泛化上界，但这些理论结果对于分析随机迭代学习算法的泛化能力具有重要意义，后续分析将对这些结果进行讨论。

4.3.2 基于条件互信息的泛化上界

Steinke 等[32]的开创性工作引入了超样本设定下的信息论泛化分析方法，其可通过全新的条件信息度量构建期望泛化误差的理论上界。此方法涉及假设 W 与超样本变量 S 之间、给定超样本数据集 $\tilde{\mathbf{Z}}$ 时的条件互信息。随后，Haghifam 等[55,58]的研究工作对此上界做了进一步改进与推广：

引理 4.5 ([58]，定理 2.6) 假设 $m=1$ 且 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意 $k\in[1,n]$ 有

$$|\overline{\mathrm{gen}}|\leq\frac{1}{|\mathsf{C}_n^k|}\sum_{u\in\mathsf{C}_n^k}\mathbb{E}_{\tilde{\mathbf{Z}}}\sqrt{\frac{2}{k}\tilde{I}^{\tilde{\boldsymbol{\pi}}}(W;S_u)}. \tag{4-9}$$

在上述引理中设 $k=n$ 即可得到 $O(\sqrt{1/n})$ 的收敛速率。将这一理论结果推广至多点学习场景将由于损失项之间的相关性而面对上述非独立同分布挑战。遵循上一小节中讨论的互信息归约泛化分析技术，本小节证明了如下定理，对于多点学习场景达到了与上述结果相同的 $O(\sqrt{m/n})$ 收敛率：

定理 4.6 假设 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意 $k\in\left[1,\frac{n}{m}\right]$ 有

$$|\overline{\mathrm{gen}}|\leq\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{\tilde{\mathbf{Z}}}\sqrt{\frac{2}{k}\tilde{I}^{\tilde{\boldsymbol{\pi}}}(W;S_u)}. \tag{4-10}$$

上述定理通过涵盖任意 $m>1$ 的多点学习情形推广了引理 4.5。注意到此类上界相较于 Steinke 等[32]工作中的初始版本在紧致性上亦有所改进，这是由于该定理将期望移至平方根运算之外，从而根据 Jensen 不等式可进一步退化至基于条件互信息 $I(W;S_u|\tilde{\mathbf{Z}})$ 的泛化上界。与定理 4.3 类似，可进一步基于经验风险 L_n 与经验和总体风险的平均值 $(L_n+L)/2$ 间的二值 KL 散度构建如下的泛化上界：

定理 4.7 假设 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意 $k\in\left[1,\frac{n}{m}\right]$ 有

$$d\left(L_n \middle\| \frac{L_n+L}{2}\right) \le \frac{1}{k|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}} I(W;S_u|\widetilde{\mathbf{Z}}). \tag{4-11}$$

条件互信息的固有特性保证了 $I(W;S_u|\widetilde{\mathbf{Z}}) \le H(S_u) = km\log 2$，从而确保了此类条件互信息泛化上界的有界性。此外，可以证明条件互信息 $I(W;S|\widetilde{\mathbf{Z}})$ 始终比输入—输出互信息 $I(W;\widetilde{\mathbf{Z}}_S)$ 上界更紧：注意到 Markov 链 $(\widetilde{\mathbf{Z}},S)-\widetilde{\mathbf{Z}}_S-W$，可得 $I(W;\widetilde{\mathbf{Z}},S|\widetilde{\mathbf{Z}}_S)=0$，继而有 $I(W;\widetilde{\mathbf{Z}}_S)=I(W;\widetilde{\mathbf{Z}},S)=I(W;S|\widetilde{\mathbf{Z}})+I(W;\widetilde{\mathbf{Z}})$。

基于上述定理 4.4 中的方法，可将相关分析拓展至解构互信息 $\widetilde{F}(W;S_u)$。通过取 $\varphi(x)=\sqrt{x}$ 或 $\varphi(x)=x$ 并随后对 $\widetilde{\mathbf{Z}}$ 取期望，可得定理 4.6 与定理 4.7 对于参数 k 均严格非递增。因此，选择 $k=1$ 即可得到最紧的泛化误差估计结果。

定理 4.8 设 $\varphi:\mathbb{R}\mapsto\mathbb{R}$ 为任意非递减凹函数，则对于任意 $k\in[1,\frac{n}{m}-1]$，有

$$\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\varphi\left(\frac{2}{k}\widetilde{F}(W;S_u)\right) \le \frac{1}{\left|\mathsf{C}_n^{km+m}\right|}\sum_{u\in\mathsf{C}_n^{km+m}}\varphi\left(\frac{2}{k+1}\widetilde{F}(W;S_u)\right). \tag{4-12}$$

与此同时，本小节进一步研究了 Hellström 等[101]工作中提出的快速收敛率上界。此类上界引入了加权泛化误差 $\overline{\mathrm{gen}}_{C_1}\triangleq L-(1+C_1)L_n$，其中 $C_1>0$ 为预定义常数。该框架催生了针对期望泛化误差的快速收敛率泛化上界，其可以达到相较于传统 $\sqrt{1/n}$ 收敛率更快的 $1/n$ 收敛率。

引理 4.9 ([101]，推论 3) 假设 $m=1$ 且 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意满足 $(C_1^2+2C_1+2)(e^{C_2}-1-C_2)-C_1C_2\le 0$ 的常数 $C_1,C_2>0$，有

$$\overline{\mathrm{gen}} \le C_1L_n+\frac{1}{n}\sum_{i=1}^{n}\frac{I(W;S_i|\widetilde{\mathbf{Z}})}{C_2}. \tag{4-13}$$

随后，本小节将此类快速率泛化上界拓展至多样本损失函数，并引入基于样本子集的互信息度量，推广了初始的个体样本互信息度量泛化结果：

定理 4.10 假设 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意 $k\in\left[1,\frac{n}{m}\right]$ 与满足 $C_1\ge-\frac{\log(2-e^{C_2})}{C_2}-1$ 的 $C_1>0$，$C_2\in(0,\log 2)$ 有

$$\overline{\mathrm{gen}} \le C_1L_n+\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{I(W;S_u|\widetilde{\mathbf{Z}})}{kC_2}. \tag{4-14}$$

进一步地，在插值模式下（即 $L_n=0$ 时）有

$$L \le \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{I(W;S_u|\widetilde{\mathbf{Z}})}{k\log 2}. \tag{4-15}$$

可见，上述定理通过扩展参数 C_1 与 C_2 的取值范围，从而增强了其泛化理论上界。图

4-1 中展示了相关参数取值范围的对比结果。可见，定理 4.10 不仅允许选择大于 $\log 2/2$ 的 C_2 值，且相较于引理 4.9，在给定 C_2 时允许选择较小的 C_1 值，从而导向了更紧的泛化上界估计。此外，通过选择 $C_2 \to \log 2$，定理 4.10 实现了向插值模式的平滑过渡。

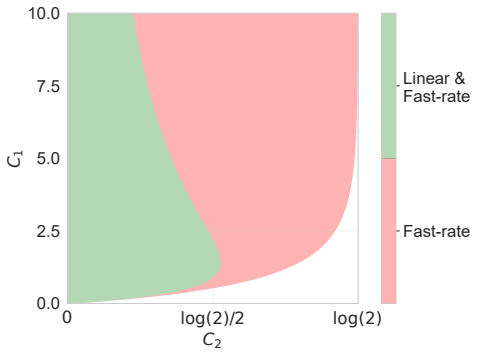
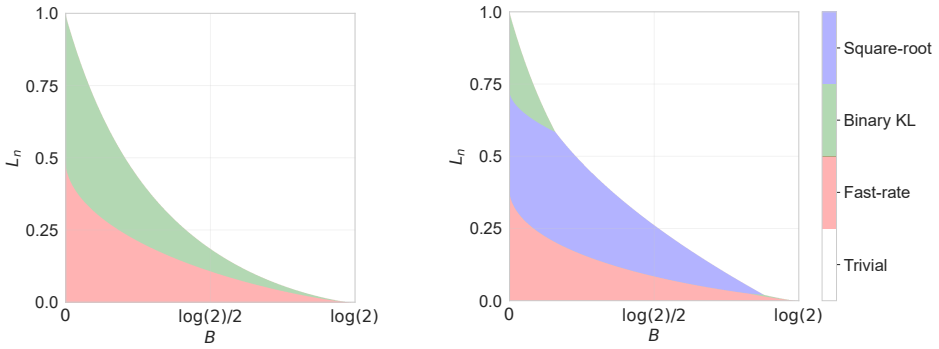


图 4-1 引理 4.9（Linear）与定理 4.10（Fast-rate）中，参数 C_1 与 C_2 的取值范围对比



(a) 常数 $\widetilde{F}(W;S_u)$ 值　　(b) 随机 $\widetilde{I^{\mathbf{Z}}}(W;S_u)$ 值

图 4-2 平方根上界（定理 4.6）、二值 KL 上界（定理 4.7）与快速率上界（定理 4.10）在取 $k=1$ 时的数值对比

本小节介绍了多种不同的泛化理论上界，但这些上界在紧致性方面的对比关系尚不明晰。为给出明确的对比结果，以下将取 $k=1$ 以考虑每个上界的最紧形式。此外，假设学习算法对于训练数据样本的排列顺序表现出期望意义下的无关性，即假设 $\widetilde{F}(W;S_u)$ 的取值与索引 u 无关。这种假设能够简化分析，且对于多数随机迭代学习算法（如 SGD）均成立。图 4-2 中将对平方根上界（定理 4.6）、二值 KL 散度上界（定理 4.7）与快速收敛率上界（定理 4.10）在不同经验风险 L_n 与互信息度量 B 取值情况下进行对比分析。图中颜色表示该情形下最紧的泛化上界，对于“Trivial”区域，没有上界能够比朴素上界 $L \leq 1$ 更紧。其中：（a）假设对于任意 $\tilde{z} \in \mathcal{Z}^{2n}$ 均有 $\widetilde{F}(W;S_u)=B$；（b）假设 $\widetilde{I^{\mathbf{Z}}}(W;S_u)$ 对于 $\widetilde{\mathbf{Z}} \sim \mu^{2n}$ 服从期望为 B 的指数分布。分析结果表明，当经验风险的值较小时，快速收敛率上界相较于其他上界表现出更优的紧致性，这种情形对于现代深度神经网络而言十分常见。与之相反，随着经验风险 L_n 的增加，二值 KL 散度上界逐渐取得优势。此外，平方根上界的紧致性取决于解构互信息 $\widetilde{I^{\mathbf{Z}}}(W;S_u)$ 的具体概率分布。在其取值较

为分散的情形下，平方根上界在 L_n 与 B 变化的中间区域表现较为突出。因此，在实际应用中应根据具体情况选择相应上界以获得最紧的泛化误差估计结果。

4.3.3 面向特定学习算法的泛化上界

本小节将以随机梯度 Langevin 动力学（SGLD）算法为例，分析多点学习场景下随机批次迭代学习算法的泛化能力。其中，SGLD 的优化轨迹可定义为 $\{W_t\}_{t=0}^{T}$，其中 $W_0 \in \mathbb{R}^d$ 代表随机初始化的模型参数向量。在第 t 步更新中，给定批次大小 b，学习算法将独立选择一组样本索引 $B_t \in [1,n]^{b\times m}$，并计算该批次中样本的平均梯度：

$$G_t = -\frac{1}{b}\sum_{u\in B_t}\nabla_w \ell(W_{t-1}, Z_u). \tag{4-16}$$

从而，SGLD 算法的更新规则可形式化为：

$$W_t = W_{t-1} + \eta_t G_t + N_t, \quad N_t \sim N(0, \sigma_t^2 I_d), \tag{4-17}$$

其中 η_t 为学习率，N_t 则为每一步迭代中加入的随机高斯噪声。

鉴于多样本损失函数的复杂性，Wang 等工作中[138]基于训练数据集独立批次分割的泛化分析技巧不再适用。从而，基于单点信息稳定性 $I(W;Z_i)$ 的泛化分析技术无法有效拓展至多点学习场景。为此，可延续第 2 章中的泛化分析技术，利用基于输入—输出互信息 $I(W;\mathbf{Z})$ 的平均信息稳定性方法推导相关学习算法的泛化上界。第 2 章中的研究表明，利用梯度协方差矩阵的行列式轨迹，可构建随机迭代学习算法输入—输出互信息的理论上界，这是一种相较于 Negrea[33,138]等工作中所探讨的梯度方差度量更为精确的泛化度量标准：

引理 4.11 (第 2 章，定理 2.13) 设 $m = 1$ 且 W_T 为 SGLD 算法 T 步更新后的参数向量，则

$$I(W;\mathbf{Z}) \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathrm{Cov}_{W_{t-1},B_t}[G_t] + I_d\right|. \tag{4-18}$$

以下进一步基于条件梯度协方差增强了这一结果，同时将其拓展至多点学习场景：

定理 4.12 设 W_T 为 SGLD 算法 T 步更新后的参数向量，则

$$I(W;\mathbf{Z}) \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d\right|, \tag{4-19}$$

其中 $\Sigma_t = \mathrm{Cov}_{B_t}[G_t]$。

根据全方差公式，条件协方差度量 Σ_t 相较于引理 4.11 中的无条件协方差严格更紧。进一步地，综合定理 4.12 与上述基于输入—输出互信息的泛化误差上界，即可得

到 SGLD 算法的泛化上界估计：

推论 4.13 假设 $\ell(\cdot,\cdot)\in[0,1]$ 且 $n \bmod m = 0$，则 SGLD 算法的总体风险满足

$$d(L_n \,\|\, L) \le \frac{m}{2n}\sum_{t=1}^{T}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d\right|. \tag{4-20}$$

虽然在此处仅关注 SGLD 算法，但通过进一步结合 Neu 等[34,141]工作中探讨的辅助优化轨迹方法，相同的泛化分析技术可拓展至推导随机梯度下降（SGD）与自适应梯度迭代（如 AdaGrad）等学习算法的泛化上界。

4.4 基于网络预测值的泛化上界

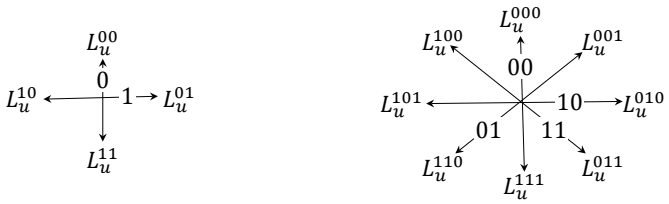
4.4.1 基于损失差异的泛化上界

Harutyunyan 等[58]的开创性工作引入了基于模型预测值与超样本变量在给定超样本数据集下的条件互信息的泛化分析技术。这一方法在 Hellström 等[51]的工作中得到了进一步完善，其聚焦于损失变量中包含的信息量，称为评估条件互信息（e-CMI）。以下是 e-CMI 上界在多点学习场景下的直接拓展：

定理 4.14 假设 $\ell(\cdot,\cdot)\in[0,1]$，则

$$|\overline{\mathrm{gen}}| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\sqrt{2I(L_u;S_u)}. \tag{4-21}$$

然而，上述互信息项的维度将随 m 呈指数级增长，从而带来新的可计算性挑战。具体而言，注意到 $|L_u| = |\mathsf{B}_m| = 2^m$ 包含了所有训练与测试样本组合对应的损失值，可得互信息 $I(L_u;S_u)$ 的维度为 $2^m + m$。Wang 等[130]工作中引入的损失差异技术可将损失项集合 L_u 替换为其差值的集合，从而将其维度减半。然而，随着 m 的增加，这一上界终将失去其在可计算性上的优势。



(a) $m = 2$ (b) $m = 3$

图 4-3 根据 $\Phi_u \in \mathsf{B}_{m-1}$ 取值选取损失对 $L_u^{\Phi_u}$ 的示意图

以下定理利用 S_{u_1} 与 Φ_u 之间的独立性，完全解决了上述可计算性问题。其上界中的关键信息度量仅涉及两个一维随机变量，从而其维度将不随 m 的增长而增长：

定理 4.15 假设 $\ell(\cdot,\cdot)\in[0,1]$，则

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(\Delta_u^{\Phi_u}; S_{u_1})}. \tag{4-22}$$

如图 4-3 所示，上述上界通过变量 Φ_u 的具体取值从损失集合 L_u 中选择一对损失值以计算 $\Delta_u^{\Phi_u}$，从而保证该互信息度量始终保持在最低维度。注意到 Markov 链 $S_{u_1} - (L_u, \Phi_u) - \Delta_u^{\Phi_u}$，利用数据处理不等式与 S_{u_1}、Φ_u 间的独立性，可证明

$$I(\Delta_u^{\Phi_u}; S_{u_1}) \leq I(L_u, \Phi_u; S_{u_1}) = I(L_u; S_{u_1}|\Phi_u) = I(L_u, S_u) - I(L_u; \Phi_u). \tag{4-23}$$

因此，定理 4.15 中的上界相较于定理 4.14 中的 e-CMI 上界严格更紧。此外，互信息项 $I(\Delta_u^{\Phi_u}; S_{u_1})$ 可理解为一个无记忆信道上可靠通信的最大速率，其输入为 S_{u_1}，输出为 $\Delta_u^{\Phi_u}$，相关详细讨论可见 Wang 等[130]的工作。这种思想可导出以下针对二值化损失函数在插值模式下的精确泛化误差刻画：

定理 4.16 假设 $\ell(\cdot,\cdot) \in [0,1]$，则在插值模式下（$L_n = 0$）有

$$L = \sum_{u \in \mathsf{P}_n^m} \frac{I(\Delta_u^{\Phi_u}; S_{u_1})}{|\mathsf{P}_n^m| \log 2} = \sum_{u \in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u}; S_{u_1})}{|\mathsf{P}_n^m| \log 2}. \tag{4-24}$$

因此，对于二值损失函数下的差值学习算法，其期望总体风险可通过 S_{u_1} 与损失对 $L_u^{\Phi_u}$ 或损失差异 $\Delta_u^{\Phi_u}$ 间的互信息之和以精确刻画。基于定理 4.6 中的相似思想，以下进一步改进了定理 4.15 中的平方根上界，通过引入解构互信息并对于 $\widetilde{\mathbf{Z}}$ 取期望：

定理 4.17 假设 $\ell(\cdot,\cdot) \in [0,1]$，则

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{\mathbf{Z}}} \sqrt{2\widetilde{I}^{\widetilde{\mathbf{Z}}}(\Delta_u^{\Phi_u}; S_{u_1})}. \tag{4-25}$$

注意到 $I(\Delta_u^{\Phi_u}; S_{u_1}) \leq I(\widetilde{\mathbf{Z}}, \Delta_u^{\Phi_u}; S_{u_1}) = I(\Delta_u^{\Phi_u}; S_{u_1}|\widetilde{\mathbf{Z}}) + I(\widetilde{\mathbf{Z}}; S_{u_1})$，进一步考虑 $\widetilde{\mathbf{Z}}$ 与 S_{u_1} 间的独立性，定理 4.15 中的互信息项 $I(\Delta_u^{\Phi_u}; S_{u_1})$ 相对于上述条件互信息 $I(\Delta_u^{\Phi_u}; S_{u_1}|\widetilde{\mathbf{Z}})$ 严格更紧。但是，若 $\widetilde{F}(\Delta_u^{\Phi_u}; S_{u_1})$ 的取值对于 $\tilde{z} \sim \mu^{2n}$ 较为分散，则定理 4.17 仍然可能导向更紧的上界估计结果，如图 4-2 所示。

4.4.2 快速收敛率的泛化上界

本小节通过进一步结合加权泛化误差以改进相关基于网络预测值的泛化上界，以提高此类上界的收敛速率。Wang 等[130]的工作中证明了如下上界：

定理 4.18 ([130]，定理 4.3) 假设 $m = 1$ 且 $\ell(\cdot,\cdot) \in [0,1]$，则存在 $C_1, C_2 > 0$ 使得

$$\overline{\mathrm{gen}} \leq C_1 L_n + \frac{1}{n} \sum_{i=1}^{n} \frac{I(L_i^0; S_i)}{C_2}. \tag{4-26}$$

在此基础上，可将相关快速收敛率上界拓展至多点学习场景，并同时考虑单点损失互信息 $2I(L_u^{\Phi_u^+};S_{u_1})$ 与损失对互信息 $I(L_u^{\Phi_u};S_{u_1})$ 间的较小值以进一步改进其紧致性。上述信息量间的差异可由交互信息 $I(L_u^{\Phi_u^+};L_u^{\Phi_u^-};S_{u_1})$ 度量，其可正可负，故不存在明确的大小关系，从而同时考虑两者将导向更紧的泛化上界：

定理 4.19 假设 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意 $C_2\in(0,\log 2)$ 与 $C_1\geq -\frac{\log(2-e^{C_2})}{C_2}-1$，有

$$\overline{\mathrm{gen}}\leq C_1 L_n+\sum_{u\in\mathsf{P}_n^m}\frac{\min\{I(L_u^{\Phi_u};S_{u_1}),2I(L_u^{\Phi_u^+};S_{u_1})\}}{|\mathsf{P}_n^m|C_2}. \tag{4-27}$$

进一步地，在插值模式下（即 $L_n=0$ 时）有

$$L\leq\sum_{u\in\mathsf{P}_n^m}\frac{\min\{I(L_u^{\Phi_u};S_{u_1}),2I(L_u^{\Phi_u^+};S_{u_1})\}}{|\mathsf{P}_n^m|\log 2}. \tag{4-28}$$

上述快速率泛化上界在经验风险接近或等于 0 时尤为高效。Wang 等[130]的工作进一步引入了经验损失方差以改进其泛化上界。受此启发，以下将其损失方差定义推广为多样本损失函数的方差：

$$V(\gamma)\triangleq\mathbb{E}_{W,\mathbf{Z}}\left[\sum_{u\in\mathsf{P}_n^m}\frac{\left(\ell(W,Z_u)-(1+\gamma)L_{\mathbf{Z}}(W)\right)^2}{|\mathsf{P}_n^m|}\right]. \tag{4-29}$$

定理 4.20 假设 $\ell(\cdot,\cdot)\in[0,1]$，则对于任意 $\gamma\in(0,1)$，$C_2\in(0,\log 2)$ 与 $C_1\geq -\frac{\log(2-e^{C_2})}{C_2\gamma^2}-\frac{1}{\gamma^2}$，

$$\overline{\mathrm{gen}}\leq C_1 V(\gamma)+\sum_{u\in\mathsf{P}_n^m}\frac{\min\{I(L_u^{\Phi_u};S_{u_1}),2I(L_u^{\Phi_u^+};S_{u_1})\}}{|\mathsf{P}_n^m|C_2}. \tag{4-30}$$

对于二值损失函数，可证明对于任意 $\gamma\in(0,1)$ 均有 $V(\gamma)=L_n-(1-\gamma^2)\mathbb{E}_{W,\mathbf{Z}}[L_{\mathbf{Z}}^2(W)]$。通过使用 $V(\gamma)$ 替代 L_n，上述损失方差上界相较于定理 4.19 对于相同的 C_1、C_2 取值至少有 $C_1(1-\gamma^2)\mathbb{E}_{W,\mathbf{Z}}[L_{\mathbf{Z}}^2(W)]$ 的提升。因此，定理 4.20 在经验风险接近但不为 0 时尤为高效。反之，当 L_n 相对较大时，图 4-2 中的分析结果表明二值 KL 散度上界能够为总体风险提供更准确的上界估计结果：

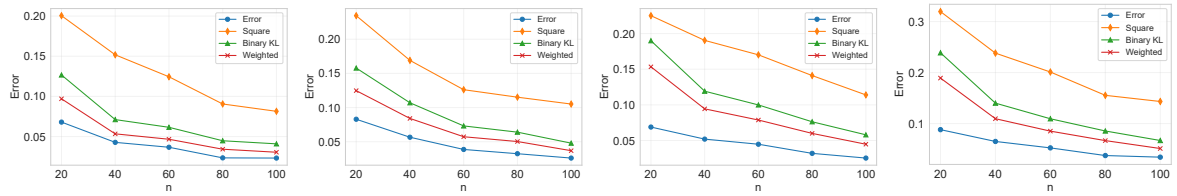
定理 4.21 假设 $\ell(\cdot,\cdot)\in[0,1]$，则

$$d\left(L_n\middle\|\frac{L_n+L}{2}\right)\leq\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}I(L_u^{\Phi_u};S_{u_1}). \tag{4-31}$$

值得注意的是，上述定理中的无条件互信息度量相较于条件互信息（即 Hellström 等[51]在单点情形下的 e-CMI 上界）严格更紧：基于 $\widetilde{\mathbf{Z}}$ 与 S_{u_1} 间的独立性，可得 $I(L_u^{\Phi_u};S_{u_1})\leq I(\widetilde{\mathbf{Z}},L_u^{\Phi_u};S_{u_1})=I(L_u^{\Phi_u};S_{u_1}|\widetilde{\mathbf{Z}})$。本小节中不同泛化上界的数值对比可参考图 4-2。

4.5 实验分析

本节将对上一节中构建的泛化上界在多种深度学习任务下进行对比评估。以下将重点对比平方根上界（定理 4.15）、二值 KL 上界（定理 4.21）与快速率上界（定理 4.19）。由于在当前图像分辨率下损失方差上界（定理 4.20）与快速率上界间的差异难以分辨，故在以下可视化结果中将其排除对比。对于以下实验，均采用二值 0-1 损失量化计算经验与总体风险。



(a) 交叉熵（$m=1$） (b) 对比损失（$m=2$） (c) 三元损失（$m=3$） (d) N-pair 损失（$m=4$）

图 4-4 模拟高斯数据集上的泛化上界对比

4.5.1 模拟数据上的泛化上界可视化

本小节的初步实验为模拟高斯数据集上的 5 分类任务。具体操作沿用了 Wang 等[130]的实验设置，通过 scikit-learn[229]库生成 5 维数据点，其中每一类别的中心点从 5 维超立方体的顶点集合中随机抽取，对应数据点则通过标准差为 0.25 的高斯分布中独立抽取。所用神经网络为 4 层 MLP 模型，其使用 ReLU 作为激活函数。损失函数的具体选择取决于 m 的值：对于 $m=1$，以下将采用二值 0-1 损失量化其泛化误差；对于 $m>1$ 的情形，则采用相应对比损失函数的二值化版本。给定网络的预测函数 $f\colon \mathcal{X}^m \to \mathbb{R}$，可通过固定阈值参数 θ 对损失值进行二值化，以双点对比损失为例：

$$L_{ij} = 1_{f(X_i,X_j)\geq\theta} \oplus 1_{Y_i=Y_j}. \tag{4-32}$$

在此，基于模型精度与召回率之间的平衡进行二分搜索以实现阈值 θ 的自适应选择。为保证统计稳定性，对于每项实验设置均进行 200 次独立实验以计算相关互信息量的估计值。如图 4-4 所示为模拟数据上的泛化上界对比结果。这些可视化结果表明，上述泛化上界均能够适应不同 m 取值下的学习场景，并能够准确反映真实泛化误差的变化趋势：其均随样本数量 n 的增加而减少。其中，本章的快速率上界表现出最强的紧致性，证实了以上在训练风险较低时的相关讨论结果。

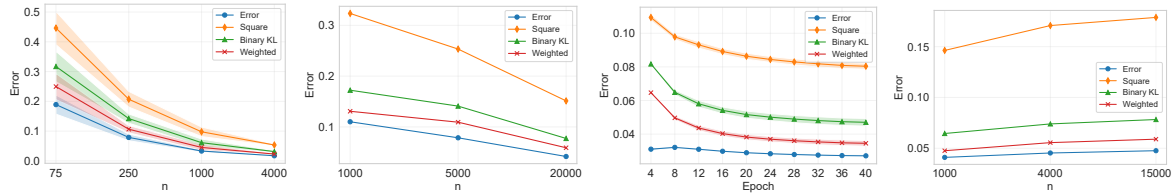
以下进一步对比了快速率上界（定理 4.19）与损失方差上界（定理 4.20）。如表 4-1 所示，在取 $\gamma=0.9$ 时，损失方差上界始终比快速率上界更紧。这验证了当经验风险接近但不为 0 时，损失方差上界所带来的优势。

表 4-1 快速率上界（左）与损失方差上界（右）间的对比结果

m \ n	20	40	60	80	100
1	0.09701 / 0.09679	0.05337 / 0.05333	0.04662 / 0.04657	0.03422 / 0.03418	0.03046 / 0.03045
2	0.12478 / 0.12452	0.08405 / 0.08380	0.05727 / 0.05712	0.05027 / 0.05010	0.03686 / 0.03681
3	0.15337 / 0.15176	0.09442 / 0.09385	0.07873 / 0.07832	0.05999 / 0.05973	0.04478 / 0.04455
4	0.18916 / 0.18782	0.10948 / 0.10895	0.08510 / 0.08475	0.06653 / 0.06622	0.05102 / 0.05085

4.5.2 真实数据上的泛化上界可视化

随后，将实验分析拓展至多个实际应用中的典型深度学习场景。遵循 Harutyunyan 等[51,58]的实验设定，以下共在四组实验下对比本章中提出的泛化上界：（1）通过 Adam 优化 MLP 网络进行 MNIST 上的二分类；（2）通过 SGLD 优化 MLP 网络进行 MNIST 上的二分类；（3）在 CIFAR-10 数据集上微调预训练的 ResNet-50 模型；（4）在 Flickr30k 数据集上微调预训练的 CLIP (ViT-B/32) 模型。



(a) CNN, Adam　(b) ResNet, SGD　(c) CNN, SGLD　(d) CLIP, Adam

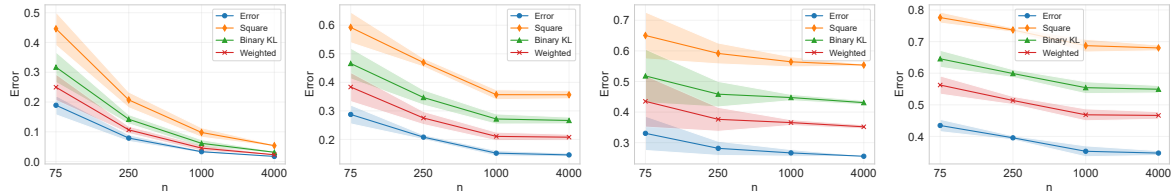
图 4-5 真实学习场景中的泛化上界对比

对于每个任务，从对应数据集中抽取 k_1 个 $\tilde{\mathbf{Z}}$ 的实例，每个实例包含随机选择的 $2n$ 个样本以构成超样本数据集。对于每个 $\tilde{\mathbf{Z}}$，继而抽取 k_2 个超样本变量 S，故共有 $k1 \times k2$ 次独立训练过程，其中 $k1$ 与 $k2$ 取值延续了 Harutyunyan 等[58]的设置。值得注意的是，对于 CLIP 模型，其经验风险或总体风险为单点与双点损失的组合。设 $\mathcal{I}$ 与 $\mathcal{T}$ 分别表示图像与文本空间，则每个样本由图像—文本对 (I_i, T_i) 构成。设 $f: \mathcal{I} \times \mathcal{T} \mapsto \mathbb{R}$ 为参数化 CLIP 模型的预测函数，则自监督对比学习的损失值可定义如下：

$$L_{ij} = \begin{cases} 1_{f(I_i,T_i)\le\theta}, & \text{if } i = j, \\ 1_{f(I_i,T_j)\ge\theta}, & \text{if } i \neq j. \end{cases} \tag{4-33}$$

通过结合 $m = 1$ 与 $m = 2$ 情形下的泛化上界，即可得到针对此类混合损失的泛化上界估计。类似地，通过平衡模型的精度与召回率以自适应选择阈值 θ。需要注意的是，预训练任务上模型的泛化性能并不与其在下游任务上的性能直接相关，尤其是在自监督学习场景下，其监督数据可能存在错误标签。因此，对于 CLIP 预训练任务，其泛化误差随样本数量 n 的增加而增加属于正常现象。真实学习任务下的泛化上界对比结果如

图 4-5 所示。可见，本章中的快速率上界（定理 4.19）始终提供了最紧的泛化误差估计结果。在所有场景中，上述泛化上界均能够良好地反映真实泛化误差的变化趋势。



(a) $\delta = 0$ (b) $\delta = 0.05$ (c) $\delta = 0.1$ (d) $\delta = 0.15$

图 4-6 随机标签噪声下的泛化上界对比

此外，为评估严重过拟合情况下泛化上界的准确度，以下在二值化 MNIST 数据集中引入随机标签噪声。具体而言，根据一定概率 δ 随机翻转数据标签。如图 4-6 所示，本章所建立的泛化上界均能够提供有效的泛化误差估计结果，其中快速率上界（定理 4.19）相较于其他上界显著更紧。

4.6 本章小结

本章克服了将现有信息论单点学习泛化上界拓展至多点学习所面临的多项重大挑战，提出了首个信息论视角下的多点学习泛化理论上界，其可适用于任意有界多点损失函数，且能够将单点、双点、三点及更高阶情形在内的常见学习场景纳入统一的泛化分析框架中。首先，目前单点学习的信息论上界依赖于损失项之间的独立同分布性质，而这种性质在多点学习场景中不复存在，带来了第一个挑战。为此，利用期望算子的线性性质将泛化误差进行拆分，继而通过互信息度量的超可加性进行上界归约，从而克服了非独立同分布挑战。其次，在超样本框架下推广现有上界将面临维度爆炸问题，其关键信息度量的维度随 m 呈指数增长，带来了第二个挑战。为此，基于超样本变量间的异或算子对其进行独立性拆分，进而通过固定维度的低维互信息构建了泛化误差上界，从而克服了维度爆炸挑战。本章后续详细讨论了上述泛化上界相较于现有上界在紧致性、可计算性以及适用范围方面的改进，从全新的信息论视角建立了统一的多点学习泛化分析框架。

本章的研究工作发表于机器学习顶级会议、CCF 推荐 A 类学术会议 International Conference on Machine Learning，论文题目为“Towards Generalization beyond Pointwise Learning: A Unified Information-theoretic Perspective”。

5 基于信息论的领域泛化理论与算法设计

5.1 引言

传统机器学习模型往往假设数据样本满足独立同分布条件，并基于此进行经验误差最小化 (Empirical Risk Minimization, ERM) 以寻找最优的模型假设。然而，在各种实际学习任务中，常常会遇到训练集与测试集数据分布不同的情形，其被称为分布偏移 (Distribution Shift) 问题。此类偏移将导致机器学习模型过拟合于训练数据分布的特定相关性，从而在面对分布外 (Out-of-distribution, OOD) 数据时，将对模型实际性能产生负面影响[230-233]。为此，国内外学者近期将一类分布外泛化问题抽象为领域泛化 (Domain Generalization) 问题：具体而言，领域泛化通过不同的领域代表不同的数据分布，并将数据集按领域划分为多个子集。其中，训练数据对应的领域称为源域 (Source Domain)，测试数据则称为目标域 (Target Domain)。基于不同领域之间共享某些不变底层相关性的基础假设，领域算法通过学习这种不变性以降低分布偏移对于模型性能的影响，使得特定的领域变化不会显著影响模型性能，从而能够在分布外测试数据上实现泛化。这种思想启发了众多不同的领域泛化算法设计，包括不变表示学习[234-235]、对抗学习[236-237]、因果推断[238-239]、梯度操纵[240-242]和鲁棒性优化[243-245]等。

现有的领域泛化理论分析多将此问题表述为一个平均情形[246-247]或最坏情形[238,243]下的优化问题。然而，优化模型的平均风险在有限源域情形下等价于经验风险最小化，其无法有效利用领域划分信息，从而对分布外数据不具备鲁棒性[238,248]。而优化模型的最大风险则受少数低概率极端数据分布情形的显著制约，其将导致对解空间的过度约束[245]。本章引入了领域泛化问题的一种新型概率刻画方式，其旨在以高概率最小化训练与测试总体风险间的差异。后续泛化理论分析表明，学习算法的输入—输出互信息与特征空间的协变量偏移在约束此泛化误差方面起到决定性作用，且可分别通过梯度分布对齐与特征分布对齐以控制。

尽管现有的相关文献已对基于分布对齐技术的领域泛化算法设计进行了广泛讨论，但此类方法通常缺乏相应的泛化理论保证，或依赖于特定强假设：包括可控的不变特征[240]、二次碗损失景观[242]或 Lipschitz 连续梯度[249]等。此类假设在现代深度学习模型中往往难以得到满足。相比之下，本章的泛化分析仅依赖于一种经过进一步松弛的领域独立同分布假设[245]，在此之上推导了适用于普适随机学习算法的泛化上界。本章的理论分析表明，通过结合梯度对齐与特征对齐，本章的新型领域泛化算法将能够有效最小化领域泛化误差。更重要的是，本章首次揭示了两者间的互补性，指出单独依靠梯度或特征对齐均无法完整解决领域泛化问题。

本章在信息论视角下给出了领域泛化问题的理论分析框架，并基于相关泛化误差上

界推导指出了影响领域泛化算法性能的关键因素，据此提出了基于域间分布对齐的新型领域泛化算法，在多个基准评估数据集上表现出最优性能。具体而言，本章突破了平均或最坏情形下的领域泛化问题表述，在经验风险最小化的基础上引入了概率泛化误差上界约束，从而构建了概率情形下的领域泛化优化目标函数。随后，将领域泛化误差进一步分解为源域与目标域上的泛化误差，并基于信息论泛化分析方法分别通过学习算法的输入—输出互信息与特征空间的协变量偏移构建了源域与目标域泛化误差的理论上界。进一步地，本章证明了输入—输出互信息可通过对齐不同源域的模型梯度分布最小化，而特征空间协变量偏移则可通过对齐域间特征分布最小化。特别地，上述分析结果揭示了两者间的互补性，从而现有工作中单独依靠梯度或特征对齐的算法均无法完整解决领域泛化问题。基于这些理论结果，进一步提出了域间分布对齐 (Inter-domain Distribution Matching, IDM) 算法，通过同时对齐源域间的梯度与特征分布以实现高概率领域泛化。此外，本章指出了传统基于低阶矩匹配的分布对齐方法面对高维与复杂概率分布时的局限性。为此，进一步提出了逐点分布对齐 (Per-sample Distribution Matching, PDM) 算法，其通过对数据点的每一维分别进行排序与对齐实现。IDM 与 PDM 算法的组合在 Colored MNIST 数据集[238]与 DomainBed 基准评估数据集[250]上达到了目前最优的综合性能表现。

总体而言，本章的主要贡献包括：（1）突破了目前平均或最坏情形下的领域泛化目标刻画，引入了概率情形下的领域泛化优化目标，重点关注学习算法以高概率最小化领域泛化误差的能力。相较于现有领域泛化理论，本章的方法显著松弛了其前提假设，并允许引入信息论方法进行泛化分析。（2）在信息论视角下分别推导了源域与目标域泛化误差的理论上界，阐明了其与梯度或特征对齐间的本质联系。特别地，本章揭示了上述两者之间的互补关系，表明任一方法均无法单独解决领域泛化问题。（3）提出了域间分布对齐（IDM）算法，首次通过结合域间梯度与特征对齐方法实现以高概率最小化领域泛化误差。本章进一步提出了基于数据点切片与排序的逐点分布对齐（PDM）算法，其与 IDM 相结合在多个领域泛化基准评估数据集上表现出最优的综合性能。

5.2 基本概念与问题设定

设 $\mathcal{W}$ 为假设空间。给定任意 $w \in \mathcal{W}$，定义 $f_w : \mathcal{X} \mapsto \mathcal{Y}$ 为对应的模型预测函数，其由编码器 $f_\varphi : \mathcal{X} \mapsto \mathcal{T}$ 与分类器 $f_\psi : \mathcal{T} \mapsto \mathcal{Y}$ 构成，其中 $\mathcal{T}$ 为特征（表示）空间。遵循 Eastwood 等[245]的工作，假设存在领域空间 $\mathcal{D}$ 上的未知分布 ν，其中每个领域 $d \in \mathcal{D}$ 均对应某个特定的数据分布 $\mu_d = P_{Z|D=d}$。在未指定领域时，$\mu = \mathbb{E}_{D\sim\nu}[\mu_d]$ 为数据的边缘分布。源域 $D_s = \{D_i\}_{i=1}^{m}$ 与目标域 $D_t = \{D_k\}_{k=1}^{m'}$ 均为从 ν 中采样的随机变量。令 $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1}^{m}$ 为训练数据集，其中每个子集 $\mathbf{Z}_i = \{Z_j^i\}_{j=1}^{n}$ 包含从数据分布 μ_{D_i} 中独立同分布采样的 n 个样本。定义领域泛化学习算法 $\mathcal{A} : \mathcal{D}^m \mapsto \mathcal{W}$，其将源域集合 D_s 作为输入

（通过代理 $\mathbf{Z}$），并依概率输出假设 $W = \mathcal{A}(D_s)$。给定损失函数 $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$，假设 $w \in \mathcal{W}$ 在特定领域 $d \in \mathcal{D}$ 上的性能表现可通过领域总体风险衡量：

$$L_d(w) = \mathbb{E}_{Z\sim\mu_d}[\ell(f_w(X), Y)]. \tag{5-1}$$

进一步地，可定义领域泛化的全局总体风险：

$$L(w) = \mathbb{E}_{D\sim v}[L_D(w)] = \mathbb{E}_{Z\sim\mu}[\ell(f_w(X), Y)]. \tag{5-2}$$

在实际应用中，领域分布 v 通常未知。因此，仅能够通过源域或目标域上的总体风险评估模型的实际泛化能力：

$$L_s(w) = \frac{1}{m}\sum_{i=1}^{m} L_{D_i}(w), \qquad L_t(w) = \frac{1}{m'}\sum_{k=1}^{m'} L_{D_k}(w). \tag{5-3}$$

以下列出了本章中用于理论分析的主要假设：

假设 5.1 源域集合 D_t 独立于目标域集合 D_s。

假设 5.2 损失函数 $\ell(\cdot,\cdot)$ 的取值范围为 $[0, M]$。

假设 5.3 给定任意 $w \in \mathcal{W}$，损失函数 $\ell(f_w(X), Y)$ 对于 $Z \sim \mu$ 满足 σ-次高斯性。

假设 5.4 损失函数 $\ell(\cdot,\cdot)$ 满足对称性与三角不等式，即对于任意 $y_1, y_2, y_3 \in \mathcal{Y}$，有 $\ell(y_1, y_2) = \ell(y_2, y_1)$ 且 $\ell(y_1, y_2) \leq \ell(y_1, y_3) + \ell(y_3, y_2)$。

假设 5.5 对于任意 $w \in \mathcal{W}$，损失函数 $\ell(f_w(X), Y)$ 对于特定度量 c 满足 β-Lipschitz 连续性，即对于任意 $z_1, z_2 \in \mathcal{Z}$，有 $|\ell(f_w(x_1), y_1) + \ell(f_w(x_2), y_2)| \leq \beta c(z_1, z_2)$。

其中，领域独立性（假设 5.1）在多数实际场景中自然满足。例如在医学图像识别任务中，设数据来源的医院代表不同的领域，则目标域应在全球所有医院中随机抽取，而源域则通过部分合作医院收集。显然，目标域与源域的抽取过程没有相关性。次高斯性（假设 5.3）是信息论泛化分析中的常见假设之一[28,33-34,141]。值得注意的是，由于有界随机变量总是满足次高斯性，假设 5.2 比假设 5.3 严格更强。Lipschitz 连续性（假设 5.5）是基于稳定性的泛化分析的关键假设，其同样被用于推导基于 Wasserstein 距离的泛化上界[64,174,177-180]。当使用距离函数（如平均绝对误差或 0-1 损失）作为损失函数时，假设 5.4 自然满足。此类假设也在现有工作中广泛采用[251-253]。

5.2.1 概率领域泛化

模型泛化的最终目标在于最小化总体风险。类似地，领域泛化的最终目标在于最小化全局总体风险：$\min_w L(w)$。在实际应用中，通常仅能够得到有限个源域下的采样数据，因此该目标退化为经验风险最小化（ERM）：$\min_w L_s(w)$。然而，近期研究发现

ERM 无法学习不同领域数据分布的不变性特征[238,248]。为此，后续研究将领域泛化描述为最坏情形下的优化问题：$\min_w \max_d L_d(w)$。然而，这在实践中将面临相同的有限源域问题，无法保证覆盖所有的边界情形[245,254]，因此通常需要对数据或特征分布的相关强假设，包括其底层因果机制的线性性质[238,244,255]或严格可分的不变与可变特征[256]等。此类假设在实际应用中往往无法满足，或将导致其解空间过分受限。本小节基于更弱的前提假设提出了以下概率优化目标：

问题 5.6 定义概率领域泛化的优化目标为：

$$\min_{\mathcal{A}} \mathbb{E}[L_s(W)], \quad s.t. \quad \mathbb{P}(|L_t(W) - L_s(W)| \geq \varepsilon) \leq \delta. \tag{5-4}$$

上述优化问题直接刻画了领域泛化的最终目标，即最优学习算法 $\mathcal{A}$ 应能够最小化源域 $L_s(W)$ 与目标域 $L_t(W)$ 总体风险间的泛化误差。上述概率取自于领域的采样过程（D_s 与 D_t）以及学习算法（W）的联合分布。值得注意的是，由于假设 5.1 允许源域（或目标域）之间存在相关性，上述独立性假设显著弱于 Eastwood 等[245]采用的独立同分布领域假设，故在实践中更加容易满足。

5.3 信息论泛化分析

领域泛化的主要目标是处理不同领域 d 所对应的数据生成分布 μ_d 不同引起的分布偏移问题，以降低其对于神经网络训练的影响，使模型在目标数据分布上具备良好的泛化能力。由分布偏移带来的数据不一致性可通过数据样本 Z 与领域 D 间的互信息 $I(Z;D)$ 来度量，该互信息可进一步分解为：

$$I(Z;D) \text{（分布偏移）} = I(X;D) \text{（协变量偏移）} + I(Y;D|X) \text{（概念偏移）}. \tag{5-5}$$

在 Federici 等[257-258]的工作中，D 为用于区分训练与测试数据样本的二值变量，而上述分析将这一概念拓展至任意离散或连续的领域空间，每一领域 $d \in \mathcal{D}$ 对应一种特定的数据分布 μ_d。上式的右侧则度量了边际输入分布 P_X（协变量偏移）以及预测分布 $P_{Y|X}$（概念偏移）在不同领域中的不一致性。以下定理表明，任意学习算法可实现的最优全局总体风险将受到概念偏移的限制：

定理 5.7 对于任意预测分布 $Q_{Y|X}$，有

$$D(P_{Y|X,D} \,\|\, Q_{Y|X}) \geq I(Y;D|X). \tag{5-6}$$

当损失函数 ℓ 为交叉熵损失时，若 $H(Y|X,D) = 0$（即标签 Y 可通过 X 与 D 完整推断），则预测器 Q 在领域 d 上的总体风险可表示为 KL 散度 $D(P_{Y|X,D=d} \,\|\, Q_{Y|X})$。这意味着一旦数据分布中存在概念偏移，则任何在训练域中拟合良好的机器学习模型总是会

在某些测试域上表现较差。这一观察验证了问题 5.6 中描述的优化与泛化之间的权衡，并强调了解决领域泛化问题的根本挑战。

5.3.1 领域泛化误差分解

本小节将展示，通过将全局总体风险 $L(W)$ 作为连接源域与目标域总体风险的桥梁，可将问题 5.6 中的约束项进一步分解为源域与目标域上的泛化误差。具体而言，由于模型 W 是通过在源域 D_s 上训练而获得，故可合理假设其在源域上的总体风险低于平均水平，即有 $L_s(W) \leq L(W)$。另外，由于测试域的采样过程与模型训练过程无关，目标域总体风险 $L_t(W)$ 可视为全局总体风险 $L(W)$ 的一个无偏估计，即有 $L(W) \approx L_t(W)$。结合上述观察结果，可合理假设 $L_s(W) \leq L(W) \approx L_t(W)$，即全局总体风险 $L(W)$ 可作为连接源域与目标域总体风险的自然桥梁。给定任意 $\lambda \in (0,1)$，可证明

$$\mathbb{P}(|L_s(W) - L_t(W)| \geq \varepsilon) \leq \mathbb{P}(|L_s(W) - L(W)| \geq \lambda\varepsilon) + \mathbb{P}(|L_t(W) - L(W)| \geq (1-\lambda)\varepsilon). \tag{5-7}$$

上式右侧的第一个事件与假设 W 高度相关，而第二个事件则与假设 W 无关。这一观察结果启发了分别探索假设无关与假设相关的泛化上界，以发现控制源域与目标域泛化误差的关键因素。

5.3.2 源域泛化误差上界

本小节首先讨论实现源域泛化的一种充分条件。受近期基于信息论的泛化分析方法启发[54,58]，以下通过领域输入—输出互信息 $I(W;D_i)$ 量化观察源域前后假设分布的变化情况，构建了如下的源域泛化误差上界：

定理 5.8 若假设 5.2 成立，则有

$$\mathbb{P}\Big(|L_s(W) - L(W)| \geq \varepsilon\Big) \leq \frac{M}{m\varepsilon\sqrt{2}} \sum_{i=1}^{m} \sqrt{I(W,D_i)} + \frac{1}{\varepsilon}\mathbb{E}_{W,D}|L_D(W) - L(W)|, \tag{5-8}$$

其中 $D \sim v$ 为独立于 W 采样的领域。

直观而言，通过提取不同领域之间共有的输入 X 与标签 Y 之间的相关性，能够增强机器学习模型的领域泛化能力。当模型 W 从特定训练域 D_i 中学习到的相关性也存在于其他训练域中时，互信息 $I(W;D_i)$ 将趋于 0。若进一步假设源域之间互相独立，则可证明每个源域对应的输入—输出互信息 $I(W,D_i)$ 之和小于总体互信息 $I(W;D_s)$，而总体互信息则度量了模型从源域中实际学习的信息量。若对于任意 $i \in [1,m]$ 均有 $I(W;D_i) = 0$，则模型实现了源域上的泛化。但这并不意味着模型没有从源域 D_s 中学习到任何信息，即依然可能有 $I(W;D_s) > 0$：

例子 5.9 设 $D_1, D_2 \in \{0,1\}$ 独立且服从 $\frac{1}{2}$-Bernoulli 分布，并设 $W = D_1 \oplus D_2$，其中 $\oplus$ 为

异或运算，则有

$$I(W;D_1)=0,\quad I(W;D_2)=0,\quad I(W;D_1,D_2)=\log 2.$$

通过在优化源域总体风险 $L_s(W)$ 的同时最小化每个 $I(W;D_i)$，则可鼓励学习算法在学习域不变相关性的同时丢弃领域特定的相关性，从而提高模型的领域泛化性能。本质上，实现源域上的泛化与减轻有监督学习中的过拟合现象类似，过拟合通常由训练数据样本不足导致，其将使模型捕捉到仅存在于训练集而不存在于测试集的输入—标签相关性[259]。类似地，源域上的泛化误差来自于源域数量不足，导致模型所捕捉的相关性在目标域中不复存在[238,260]。

上述定理 5.8 中，上界右侧的第二项与目标域泛化误差高度相关，将在下一小节中进行展开分析。

下面，将展示最小化互信息 $I(W;D_i)$ 可通过对齐不同源域中的条件梯度分布实现。为此，假设学习算法 $\mathcal{A}$ 为随机迭代优化算法，例如随机梯度下降（SGD）。在第 t 步迭代中，其模型参数的更新规则可表述为：

$$W_t=W_{t-1}-\eta_t\sum_{i=1}^{m}g(W_{t-1},B_t^i),\qquad g(w,B_t^i)=\frac{1}{m|B_t^i|}\sum_{z\in B_t^i}\nabla_w\ell(f_w(x),y),\tag{5-9}$$

其中 W_0 为初始化参数向量，η_t 为学习率，B_t^i 为从源域 D_i 中随机抽取的数据样本集，用于计算迭代梯度。假设算法 $\mathcal{A}$ 共进行 T 步迭代，则有：

定理 5.10 设 $G_t=-\eta_t\sum_{i=1}^{m}g(W_{t-1},B_t^i)$，则有

$$I(W_T;D_i)\le\sum_{t=1}^{T}I(G_t;D_i|W_{t-1}).\tag{5-10}$$

虽然上述分析基于 SGD 的更新规则进行，但最终结论同样适用于各种随机迭代学习算法，例如 SGLD 与 AdaGrad 算法。上述定理表明，通过在每一步迭代过程中最小化互信息 $I(G_t;D_i|W_{t-1})$，即可最小化领域输入—输出互信息 $I(W_T;D_i)$，从而实现源域上的泛化。值得注意的是，条件互信息 $I(G_t;D_i|W_{t-1})$ 可重写为 KL 散度 $D(P_{G_t|D_i,W_{t-1}}\,\|\,P_{G_t|W_{t-1}})$，从而直接启发对齐不同源域上的梯度分布：

命题 5.11 若 $\mathbb{E}_{W,D}|L_D(W)-L(W)|\to 0$，则域间梯度对齐可最小化源域泛化误差。

直观而言，梯度对齐将强制模型学习不同源域间共享的输入—标签相关性，从而防止过拟合至领域特定的相关性，从而改善模型的领域泛化性能[241-242]。

以下进一步提出一种替代上界推导方法，其基于损失函数的 Lipschitz 连续性而非次高斯性，可导向相较于互信息度量而言更紧的泛化误差上界：

定理 5.12 若损失函数 $\ell(f_w(X),Y)$ 对于 w 满足 β'-Lipschitz 连续性，则有

$$\left|\mathbb{E}_{W,D_s}[L_s(W)] - \mathbb{E}_W[L(W)]\right| \leq \frac{\beta'}{m}\sum_{i=1}^{m}\mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)]. \tag{5-11}$$

相较于 KL 散度度量，Wasserstein 距离不仅满足对称性，且可导向更紧的泛化误差上界：假设其所使用的度量 c 为离散度量，则有以下归约：

$$\mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)] = \mathbb{E}_{D_i}[\mathrm{TV}(P_{W|D_i}, P_W)] \leq \mathbb{E}_{D_i}\sqrt{\frac{1}{2}D(P_{W|D_i} \,\|\, P_W)} \leq \sqrt{\frac{1}{2}I(W;D_i)}, \tag{5-12}$$

其中 TV 为全变差。这一观察证实了输入—输出互信息 $I(W;D_i)$ 也可作为其他替代泛化度量（即全变差或 Wasserstein 距离）的理论上界。因此，通过最小化互信息 $I(W;D_i)$，不仅对于优化过程而言更加稳定[253,261]，且可同时优化其余替代度量。

5.3.3 目标域泛化误差上界

下面，本小节将讨论目标域泛化的一种充分条件。由于模型训练过程独立于目标域的采样过程，因此可视作常量 $w \in \mathcal{W}$。由于目标域的分布 v 相同，容易验证 $\mathbb{E}_{D_t}[L_t(w)] = L(w)$。随后，可建立如下的目标域泛化误差上界：

定理 5.13 若假设 5.3 成立，则对于任意 $w \in \mathcal{W}$，有

$$\mathbb{P}(|L_t(w) - L(w)| \geq \varepsilon) \leq \frac{\sigma}{\varepsilon}\sqrt{2I(Z;D)}. \tag{5-13}$$

上述理论结果可从两方面进行解读：首先，在随机抽样的测试域中评估模型 w 可反映其全局总体风险，因为 $L_t(w)$ 是 $L(w)$ 的无偏估计。其次，$L(w)$ 的估计值可用于预测模型 w 在未知领域上的泛化能力，故可与定理 5.8 结合以解决问题 5.6。

在上述定理 5.13 中，模型在目标域上实现泛化的概率主要由分布偏移 $I(Z;D)$ 控制。值得注意的是，互信息 $I(Z;D)$ 是数据收集过程的内在属性，因而无法从学习算法的角度进行优化。为此，可将编码器 φ 视为数据预处理过程的一部分，并通过学习算法能够优化的特征空间分布偏移构建泛化误差上界。在与定理 5.13 相同的条件下，对于任意分类器 ψ，有：

$$\mathbb{P}(|L_t(\psi) - L(\psi)| \geq \varepsilon) \leq \frac{\sigma}{\varepsilon}\sqrt{2I(T,Y;D)}. \tag{5-14}$$

类似地，特征空间的分布偏移存在以下分解：

$$I(T,Y;D)\ \text{（分布偏移）} = I(T;D)\ \text{（协变量偏移）} + I(Y;D|T)\ \text{（概念偏移）}. \tag{5-15}$$

这一观察结果启发了通过同时最小化特征空间的协变量偏移与概念偏移以实现目标域上的泛化。以下进一步展示，通过单独最小化协变量偏移 $I(T;D)$，便足以约束目标域泛化误差：

定理 5.14 若假设 5.2 与假设 5.4 成立，则对于任意 ψ，有

$$\mathbb{P}(L_t(\psi) - L(\psi) \geq \varepsilon) \leq \frac{M}{\varepsilon\sqrt{2}}\sqrt{I(T;D)} + \frac{2}{\varepsilon}L^*, \tag{5-16}$$

其中 $L^* = \min_{f^*:\mathcal{T}\mapsto\mathcal{Y}}[L(f^*)]$。

上述定理表明，模型在目标域上的泛化误差主要由协变量偏移所控制。值得注意的是，定理中的最优分类器 f^* 可从整个特征映射空间 $\mathcal{T} \mapsto \mathcal{Y}$ 中选择，而并不局限于由神经网络构成的分类器结构。在无标签噪声情形下，应存在真实标签函数 h^*，使得 $Y = h^*(T)$，此时可得 $L^* = 0$。因此，L^* 可作为真实数据分布 μ 中标签噪声水平的度量指标。此外，特征空间的协变量偏移 $I(T;D)$ 等价于 KL 散度 $D(P_{T|D} \,\|\, P_T)$，由此启发对齐不同领域的特征分布：

命题 5.15 若标签噪声水平 L^* 较低，则域间特征对齐可最小化目标域泛化误差。

在定理 5.14 的证明中，得到了如下的目标域泛化误差绝对值期望上界：

$$\mathbb{E}_{W,D}|L_D(W) - L(W)| \leq \frac{M}{\sqrt{2}}\sqrt{I(T;D)} + 2L^*. \tag{5-17}$$

这一结果补足了定理 5.8 中的源域泛化上界，证实了同时最小化 $I(W;D_i)$ 与 $I(T;D)$ 是实现源域泛化的一种充分条件。

虽然最小化特征空间协变量偏移 $I(T;D)$ 能够提高目标域泛化能力，但直接优化 $I(T;D)$ 需要目标域的数据样本，而在领域泛化的经典设定中，目标域在整个模型训练阶段均不可见。一种替代方法是转而对齐源域中的特征分布：设 $T_i = f_\varphi(X)$，其中 $(X, Y) \sim \mu_{D_i}$，则可通过最小化 $I(T_i;D_i)$ 作为最小化 $I(T;D)$ 的替代方法。以下定理验证了这种方法的可行性：

定理 5.16 在弱假设条件下，有

$$D_s(P_{T,D} \,\|\, P_{T_i,D_i}) = O\left(\sqrt{I(W;D_i)}\right), \tag{5-18}$$

其中 $D_s(P \,\|\, Q) = D(P \,\|\, Q) + D(Q \,\|\, P)$。

上述定理表明，目标域的特征联合分布 $P_{T,D}$ 与源域特征联合分布 P_{T_i,D_i} 间的差异性可通过输入—输出互信息 $I(W;D_i)$ 作为上界。通过同时最小化 $I(W;D_i)$，即可使用 $I(T_i;D_i)$ 作为最小化 $I(T;D)$ 的代理，从而实现模型在目标域上的泛化。

虽然上述分析不需要源域或目标域之间的独立性条件，但此类条件在实际应用中往往能够得到满足，并可导向更紧的泛化上界。具体而言，若目标域满足独立同分布条件，则定理 5.13 与定理 5.14 中的上界可进一步缩紧 m' 倍。

5.4 域间分布对齐算法

受上述理论分析启发，本节提出了基于域间分布对齐（IDM）的领域泛化算法。由于全局总体风险 $L(W)$ 可作为连接源域 $L_s(W)$ 与目标域 $L_t(W)$ 总体误差的自然桥梁，问题 5.6 中的正则化项可通过联合定理 5.8 与定理 5.14 中的上界，从而得到最终优化目标。具体而言，对于任意 $\lambda \in (0,1)$，有

$$\mathbb{P}(|L_t(W) - L_s(W)| \geq \varepsilon) \leq \frac{M}{m\varepsilon\lambda\sqrt{2}} \sum_{i=1}^{m} \sqrt{I(W, D_i)} + \frac{1}{\varepsilon\lambda(1-\lambda)} \left(\frac{M}{\sqrt{2}} \sqrt{I(T; D)} + 2L^* \right). \tag{5-19}$$

上述观察结果启发了同时对齐梯度与特征分布的领域泛化算法。虽然基于分布对齐的领域泛化算法设计在现有工作中已得到了广泛探索，但本文首次探索了梯度与特征对齐的互补关系：

命题 5.17 梯度与特征分布对齐是最小化问题 5.6 中约束项 $\mathbb{P}(|L_t(W) - L_s(W)| \geq \varepsilon)$ 的一种充分条件。

具体而言，实现源域泛化除对齐梯度分布外还需同时最小化目标域泛化误差的绝对值期望（定理 5.8），而实现目标域泛化除对齐源域特征分布外还需同时最小化输入—输出互信息（定理 5.16）。因此，现有仅单独关注梯度或特征对齐的工作不足以完整解决领域泛化问题。本章首次证明了梯度与特征对齐的组合是最小化领域泛化误差的一种充分条件。

5.4.1 逐点分布对齐方法

虽然现有文献中已探索了多种分布对齐方法，但此类方法往往难以处理高维复杂分布情形。一般而言，梯度或特征的具体分布在实际应用中未知，因此仅能够通过批次数据样本进行对齐。以下针对有限样本情形下的高维分布对齐给出了不可能定理：

定理 5.18 设 n 与 b 分别为用于分布对齐的数据点的维度与数量。若 $n > b + 1$，则对于任意给定的一组数据点集，存在无限多个无法由此数据点集做出区分的领域。若 $n > 2b + 1$，则存在无限多个领域，其无法区分任意给定的两组数据点集。

附录 A.5 中提供了该定理的正式表述。在真实学习场景中，特征或梯度的维度常常超过训练时的数据批次大小。由上述定理可知，在此类情形下对齐整个分布从理论上不可行，从而导致致力于对齐完整分布的方法（如 CORAL[234]或 MMD[235]）较为低效。这一观察结果得到了 Rame 等[242]工作的验证，其通过实验证实了对齐完整的梯度协方差矩阵相较于仅对齐对角线元素而言并不能达到更优秀的性能表现。此外，现有分布对齐方法集中于对齐梯度或特征的方向[241,249,262]或低阶矩[234,240,242]，此类方法并不足以应对较为复杂的分布情况。例如，虽然标准正态分布 $N(0,1)$ 与均匀分布 $U(-\sqrt{3}, \sqrt{3})$

具有相同的期望与方差，但其本质上完全不同。为此，本小节提出了逐点分布对齐技术，通过最小化概率密度估计子间的 KL 散度上界，实现逐维的分布对齐。

设 $\{x_i^1\}_{i=1}^b$ 与 $\{x_i^2\}_{i=1}^b$ 分别为从概率分布 P 与 Q 中采样得到的两组一维数据点。设 p_i 表示以 x_i^1 为期望，以 σ^2 为方差的高斯分布的概率密度函数，则 P 基于样本的核概率密度估计子可写作 $\bar{p}(x)=\frac{1}{b}\sum_i p_i(x)$（$q_i$，$\bar{Q}$ 与 $\bar{q}$ 类似）。以下定理给出了概率密度估计子之间的 KL 散度（或 Wasserstein 距离）的一种可计算上界估计：

定理 5.19 设 f 为 $[1,b]\leftrightarrow[1,b]$ 的双射，P_i 为由 p_i 定义的概率分布（Q_i 与 q_i 类似），则有 $D(\bar{P}\,\|\,\bar{Q})\le\frac{1}{b}\sum_{i=1}^b D(P_i\,\|\,Q_{f(i)})$，且 $\mathbb{W}(\bar{P},\bar{Q})\le\frac{1}{b}\sum_{i=1}^b\mathbb{W}(P_i,Q_{f(i)})$。

因此，可通过最小化数据点对应高斯分布之间的 KL 散度（或 Wasserstein 距离）实现分布对齐，其可进一步通过对齐两两采样数据点实现。以下定理给出了上述对齐顺序 f 的最优选择：

定理 5.20 设 $\{x_i^1\}_{i=1}^b$ 与 $\{x_i^2\}_{i=1}^b$ 均按相同顺序排序，则当 $f(j)=j$ 时，$\sum_{i=1}^b D(P_i\,\|\,Q_{f(i)})$ 与 $\sum_{i=1}^b\mathbb{W}(P_i,Q_{f(i)})$ 取最小值。

总体而言，PDM 的实现步骤是将数据点切分为不同维度，将每一维分别升序（或降序）排序，随后对齐来自于不同源域的排序数据点。PDM 可视为基于矩对齐的分布对齐技术的一种拓展，其通过排序并对齐的实现方法，可同时对概率分布的多阶矩进行对齐，从而解决了现有方法不足以对齐复杂分布的问题。同时，PDM 通过将数据点按维度切分，避免了无效的高维分布对齐，实现了高效的分布对齐算法。

值得注意的是，这种逐维排序对齐方案与切片 Wasserstein 距离的计算类似[263-265]。这进一步验证了 PDM 方法的有效性，表明其适用范围并不局限于高斯概率密度估计子。然而，需要注意的是，切片 Wasserstein 距离并不能直接应用于问题 5.6 的求解：在损失函数不满足 Lipschitz 连续性时，基于 Wasserstein 距离的泛化误差上界并不成立。相比之下，上述基于 KL 散度的分析过程仅需损失函数满足次高斯性。

5.4.2 算法设计

基于上述分析，于此提出域间分布对齐（IDM）算法，通过同时对齐域间梯度与特征分布实现高概率的领域泛化。回顾问题 5.6 中的附加正则化项，可通过以下 Lagrange 乘子构建 IDM 的优化目标函数：

$$\mathcal{L}_{\mathrm{IDM}}=\mathcal{L}_{\mathrm{E}}+\lambda_1\mathcal{L}_{\mathrm{G}}+\lambda_2\mathcal{L}_{\mathrm{T}}=\frac{1}{m}\sum_{i=1}^m[L_{D_i}(W)+\lambda_1\mathrm{PDM}(G_i)+\lambda_2\mathrm{PDM}(T_i)]. \tag{5-20}$$

上述优化目标中，$\mathcal{L}_{\mathrm{E}}$ 为经验风险最小化（ERM）的损失项，$\mathcal{L}$G 与 $\mathcal{L}_{\mathrm{T}}$ 分别为梯度对齐与特征对齐的损失项，其可通过上述的 PDM 分布对齐方法实现。为配合特征分布对齐，可将分类器 ψ 视为真正的预测器，同时出于对所需内存与训练时间的相关考虑，以下

仅对分类器 ψ 实施梯度对齐，这一简化方案也为 Rame 等[242]的工作所采纳。此外，超参 λ_1 与 λ_2 的取值应根据协变量偏移以及概念偏移的程度选择：首先，注意到 Markov 链 $D \to X \to T$，可得 $I(X;D)$ 是 $I(T;D)$ 的上界。因此在 $I(X;D)=0$ 时，特征分布已经自然对齐，无需额外的特征对齐操作。此外，若 $I(Y;D|X)=0$，则无需实施梯度对齐，因为此时特征对齐已足以最小化整体分布偏移。因此，λ_1 与 λ_2 的取值应分别对应于协变量偏移与概念偏移的大小。

算法 5-1 PDM for distribution matching

Input: Data matrices $\{X^i\}_{i=1}^m$, moving average γ.
Output: Penalty of distribution matching.
for $i \leftarrow 1$ **to** m **do**
  Sort the elements of X^i in each column in ascending order;
  Calculate moving average $X^i_{ma} = \gamma X^i_{ma} + (1-\gamma)X^i$;
end
Calculate the mean of data points across domains: $X_{ma} = \frac{1}{m}\sum_{i=1}^m X^i_{ma}$;
return $\mathcal{L}_{\text{PDM}} = \frac{1}{mdb}\sum_{i=1}^m \|X_{ma} - X^i_{ma}\|_F^2$.

以下给出了 PDM（算法 5-1）与 IDM（算法 5-2）算法的伪代码实现。用于分布对齐的输入数据点表示为 $X^i \in \mathbb{R}^{b\times d}$，其中 b 为批次大小，d 为数据点维度，即 X 的每一行代表一个数据点。本小节沿用了 Rame 等[242]的实现技巧，通过滑动平均操作改善概率密度估计的准确度，其可视为增大了批次的等价大小。其中，滑动平均 X^i_{ma} 初始值设为 0。值得注意的是，定理 5.18 中的分析结果仍然有效：即使采用滑动平均，该批次的等价大小（640）仍然显著小于特征（2048）或梯度（$\gg$2048）的维度，从而满足 $d > 2b+1$。

5.5 实验分析

本节在 Colored MNIST 数据集[238]与 DomainBed 基准评估数据集[250]上测试了本章所提出的 IDM 算法，以验证其在面对不同分布偏移情形时的分布外泛化能力。

5.5.1 Colored MNIST 数据集

Colored MNIST 数据集是由 Arjovsky 等[238]引入的二分类任务，其与原始 MNIST 数据集的主要区别在于手动引入了标签与图像颜色间的强相关性。Colored MNIST 通过以下过程生成：

（1） 根据数字是否大于 4 赋予样本初始标签（数字 0-4 标签为 0，数字 5-9 标签为 1）。
（2） 以 0.25 的概率随机翻转标签，因此仅通过数字做出判断的预测器最高可达 75% 准确率。

算法 5-2 IDM for high-probability domain generalization

Input: Model W, training dataset $\mathbf{Z}$, hyper-parameters $\lambda_1, \lambda_2, t_1, t_2, \gamma_1, \gamma_2$.

for $t \leftarrow 1$ **to** *#steps* **do**
  for $i \leftarrow 1$ **to** m **do**
    Randomly sample a batch $B_t^i = (X_t^i, Y_t^i)$ from $\mathbf{Z}_i$ of size b;
    Compute individual representations: $(T_t^i)_j = f_\Phi\left((X_t^i)_j\right)$, for $j \in [1, b]$;
    Compute individual risks: $(L_t^i)_j = \ell\left(f_\Psi\left((T_t^i)_j\right), (Y_t^i)_j\right)$, for $j \in [1, b]$;
    Compute individual gradients: $(G_t^i)_j = \nabla_\Psi (L_t^i)_j$, for $j \in [1, b]$;
  end
  Compute total empirical risk: $\mathcal{L}_{\text{IDM}} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (L_t^i)_j$;
  if $t \geq t_1$ **then**
    Compute gradient alignment risk: $\mathcal{L}_{\text{G}} = \text{PDM}(\{G_t^i\}_{i=1}^m, \gamma_1)$;
    $\mathcal{L}_{\text{IDM}} = \mathcal{L}_{\text{IDM}} + \lambda_1 \mathcal{L}_{\text{G}}$;
  end
  if $t \geq t_2$ **then**
    Compute representation alignment risk: $\mathcal{L}_{\text{T}} = \text{PDM}(\{T_t^i\}_{i=1}^m, \gamma_2)$;
    $\mathcal{L}_{\text{IDM}} = \mathcal{L}_{\text{IDM}} + \lambda_2 \mathcal{L}_{\text{T}}$;
  end
  Back-propagate gradients $\nabla_W \mathcal{L}_{\text{IDM}}$ and update the model W;
end

（3） 对于每个领域，分配概率 P_e 作为标签与颜色间的相关性：对于标签为 0 的样本，以概率 P_e 将其染为红色，否则染为绿色；对于标签为 1 的样本，则以概率 P_e 将其染为绿色，否则染为红色。

其中，Colored MNIST 数据集包含两个源域 $D_s = \{P_1 = 90\%, P_2 = 80\%\}$，一个目标域 $D_t = \{P_3 = 10\%\}$。这使得对于源域数据而言，仅根据图像颜色做出判断将比根据数字做出判断达到更高的准确率，而在目标域中，标签与颜色间的相关性将被反转。从而，经典的经验误差最小化（ERM）方法将过拟合至图像颜色，因此在目标域上的泛化性能较差。因此，Colored MNIST 数据集是评估领域泛化算法能否通过不同源域数据学习其不变特征的一种理想方案。

沿用 Arjovsky 等[238]的实验设置，以下采用基于 ReLU 激活函数的 3 层 MLP 网络进行训练，通过全批次梯度下降方法进行 500 次迭代。其中，超参数将通过 50 次独立试验随机搜索。本小节采用两阶段训练方法，即赋予分布对齐损失较低的初始系数 λ，并在之后的训练过程中逐步提高。基于 IDM 的训练误差曲线，以下可视化了相关领域泛化算法损失项（包括 IRM、V-Rex、IGA 和 Fishr）的动态变化情况，为了直观比较而对损失值进行了归一化操作，如图 5-1 所示。这一观察结果验证了定理 5.8 中的结果，即梯度对齐可最小化不同源域对应风险的差异，确保预测器在不同源域中保持其最优

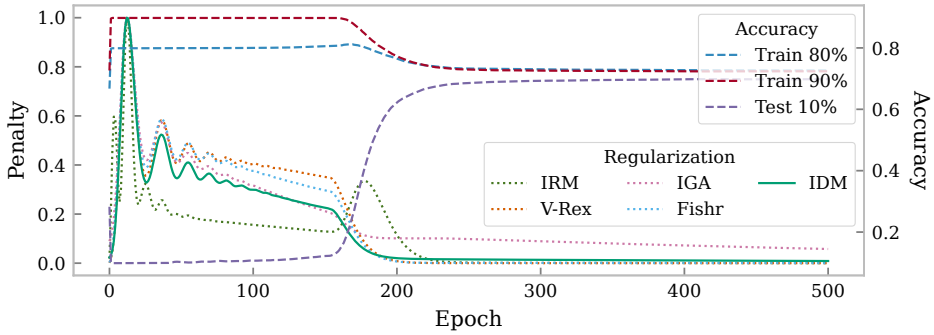


图 5-1 Colored MNIST 数据集中，总体风险与不同损失项的可视化结果

性，从而促进源域上的泛化。此外，可观察到 IDM 能够同时优化多种领域泛化损失，进一步验证了此算法的优越性。

表 5-1 Colored MNIST 数据集上不同领域泛化算法的性能对比

算法	训练集准确率	测试集准确率	灰度图准确率
ERM	86.4 ± 0.2	14.0 ± 0.7	71.0 ± 0.7
IRM	71.0 ± 0.5	65.6 ± 1.8	66.1 ± 0.2
V-REx	71.7 ± 1.5	67.2 ± 1.5	68.6 ± 2.2
IGA	68.9 ± 3.0	67.7 ± 2.9	67.5 ± 2.7
Fishr	69.6 ± 0.9	71.2 ± 1.1	70.2 ± 0.7
IDM	70.2 ± 1.4	70.6 ± 0.9	70.5 ± 0.7

表 5-1 中展示了 Colored MNIST 数据集上 10 次独立测试后的总体性能对比。沿用 Arjovsky 等[238]的实验设定，基于 $\max_w \min(L_s(w), L_t(w))$ 准则选择最优模型。可见，本章的 IDM 算法在源域与目标域性能之间实现了最优权衡（70.2%），并在灰度图像上达到了近似最优的性能表现（70.5%），仅次于 Oracle 预测器（71.0%，通过 ERM 方法在灰度图像上训练得到的模型）。

5.5.2 DomainBed 基准评估数据集

DomainBed 基准评估数据集[250]由多个模拟与真实数据集构成，可用于评估领域迁移（Domain Adaptation）或领域泛化算法的综合性能表现。为确保公平对比，DomainBed 中规定了统一的超参数调优框架，为每一超参分别设定了取值范围，并将超参数尝试次数限定为 20 次，最终结果取 3 次独立实验的均值与方差。因此，DomainBed 可作为评估领域泛化算法的一种全面且严格的测试基准。在此将 IDM 与 20 种现有方法进行对比以确保评估的全面性，总体结果可见表 5-2。

如表所示，IDM 算法在 CMNIST 数据集上达到了最优准确率（72.0%），十分接近理论最优值（75.0%），超越了目前所有基于对齐方向（AND-mask、SAND-mask、Fish）或低

表 5-2 DomainBed 数据集上不同领域泛化算法的性能对比。以下标出了**最优**，次优 与未超越 ERM的性能结果

算法	准确率 (↑)								排名 (↓)		
	CMNIST	**RMNIST**	**VLCS**	**PACS**	**OffHome**	**TerraInc**	**DomNet**	**Avg**	均值	中位数	最差
ERM	57.8 ± 0.2	97.8 ± 0.1	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	53.0 ± 0.3	41.3 ± 0.1	68.7	12.3	11	20
IRM	67.7 ± 1.2	97.5 ± 0.2	76.9 ± 0.6	84.5 ± 1.1	63.0 ± 2.7	50.5 ± 0.7	28.0 ± 5.1	66.9	18.3	20	22
GroupDRO	61.1 ± 0.9	97.9 ± 0.1	77.4 ± 0.5	87.1 ± 0.1	66.2 ± 0.6	52.4 ± 0.1	33.4 ± 0.3	67.9	11.7	10	19
Mixup	58.4 ± 0.2	98.0 ± 0.1	78.1 ± 0.3	86.8 ± 0.3	68.0 ± 0.2	**54.4** ± 0.3	39.6 ± 0.1	69.0	7.3	6	15
MLDG	58.2 ± 0.4	97.8 ± 0.1	77.5 ± 0.1	86.8 ± 0.4	66.6 ± 0.3	52.0 ± 0.1	41.6 ± 0.1	68.7	12.6	13	18
CORAL	58.6 ± 0.5	98.0 ± 0.0	77.7 ± 0.2	87.1 ± 0.5	**68.4** ± 0.2	52.8 ± 0.2	41.8 ± 0.1	69.2	6.4	5	14
MMD	63.3 ± 1.3	98.0 ± 0.1	77.9 ± 0.1	87.2 ± 0.1	66.2 ± 0.3	52.0 ± 0.4	23.5 ± 9.4	66.9	10.0	10	22
DANN	57.0 ± 1.0	97.9 ± 0.1	79.7 ± 0.5	85.2 ± 0.2	65.3 ± 0.8	50.6 ± 0.4	38.3 ± 0.1	67.7	15.0	18	22
CDANN	59.5 ± 2.0	97.9 ± 0.0	**79.9** ± 0.2	85.8 ± 0.8	65.3 ± 0.5	50.8 ± 0.6	38.5 ± 0.2	68.2	12.4	14	18
MTL	57.6 ± 0.3	97.9 ± 0.1	77.7 ± 0.5	86.7 ± 0.2	66.5 ± 0.4	52.2 ± 0.4	40.8 ± 0.1	68.5	11.7	10	21
SagNet	58.2 ± 0.3	97.9 ± 0.0	77.6 ± 0.1	86.4 ± 0.4	67.5 ± 0.2	52.5 ± 0.4	40.8 ± 0.2	68.7	11.3	9	17
ARM	63.2 ± 0.7	**98.1** ± 0.1	77.8 ± 0.3	85.8 ± 0.2	64.8 ± 0.4	51.2 ± 0.5	36.0 ± 0.2	68.1	13.0	16	21
VREx	67.0 ± 1.3	97.9 ± 0.1	78.1 ± 0.2	87.2 ± 0.6	65.7 ± 0.3	51.4 ± 0.5	30.1 ± 3.7	68.2	10.6	8	20
RSC	58.5 ± 0.5	97.6 ± 0.1	77.8 ± 0.6	86.2 ± 0.5	66.5 ± 0.6	52.1 ± 0.2	38.9 ± 0.6	68.2	13.4	13	19
AND-mask	58.6 ± 0.4	97.5 ± 0.0	76.4 ± 0.4	86.4 ± 0.4	66.1 ± 0.2	49.8 ± 0.4	37.9 ± 0.6	67.5	17.0	16	22
SAND-mask	62.3 ± 1.0	97.4 ± 0.1	76.2 ± 0.5	85.9 ± 0.4	65.9 ± 0.5	50.2 ± 0.1	32.2 ± 0.6	67.2	17.9	19	22
Fish	61.8 ± 0.8	97.9 ± 0.1	77.8 ± 0.6	85.8 ± 0.6	66.0 ± 2.9	50.8 ± 0.4	**43.4** ± 0.3	69.1	11.3	11	18
Fishr	68.8 ± 1.4	97.8 ± 0.1	78.2 ± 0.2	86.9 ± 0.2	68.2 ± 0.2	53.6 ± 0.4	41.8 ± 0.2	70.8	5.4	**3**	16
SelfReg	58.0 ± 0.7	**98.1** ± 0.7	78.2 ± 0.1	**87.7** ± 0.1	68.1 ± 0.3	52.8 ± 0.9	43.1 ± 0.1	69.4	5.0	**3**	19
CausIRL$_{\text{CORAL}}$	58.4 ± 0.3	98.0 ± 0.1	78.2 ± 0.1	87.6 ± 0.1	67.7 ± 0.2	53.4 ± 0.4	42.1 ± 0.1	69.4	5.0	**3**	15
CausIRL$_{\text{MMD}}$	63.7 ± 0.8	97.9 ± 0.1	78.1 ± 0.1	86.6 ± 0.7	65.2 ± 0.6	52.2 ± 0.3	40.6 ± 0.2	69.2	10.4	10	20
IDM	**72.0** ± 1.0	98.0 ± 0.1	78.1 ± 0.4	87.6 ± 0.3	68.3 ± 0.2	52.8 ± 0.5	41.8 ± 0.2	**71.2**	**3.3**	**3**	**6**

阶矩（Fishr）的分布对齐领域泛化算法。这验证了 PDM 分布对齐方法的优越性以及梯度与特征对齐之间的互补关系。与之相反的是，仅对齐特征分布的相关算法（CORAL、MMD、DANN、CDANN）无法解决概念偏移问题，因而在 CMNIST 数据集上表现较差。此外，IDM 算法在 RMNIST 与 PACS 数据集上达到了所有基于分布对齐的领域泛化算法之中的最优准确率，在 RMNIST（98.0% 对 98.1%）、PACS（87.6% 对 87.7%）、OfficeHome（68.3% 对 68.4%）数据集上取得了与最优算法相当的性能表现，并在所有数据集上取得了最优的平均准确率（71.2%）与最高的排名结果（包括平均、中位数以及最差排名）。此外，IDM 是唯一一个在所有数据集上均名列前茅的算法（22 种算法中的前 6 名），而其余所有方法均至少在一个数据集上落后于多数其他算法。IDM 的计算效率同样十分优秀，相较于朴素的 ERM 方法在最大数据集 DomainNet 上的训练时间仅增加了 5%，在较小数据集上的额外开销可忽略不计，如表 5-3 所示。

虽然 IDM 算法的整体表现十分优秀，但在 TerraIncognita 数据集上的性能比较落后。这可能由多种原因共同导致：首先，IDM 相较于其他算法使用了更多超参数，而 DomainBed 硬性规定了超参数调优次数，因而对 IDM 算法十分不利。上一节中的讨论结果表明，λ_1 与 λ_2 应根据协变量与概念漂移的具体程度进行选择。注意到 CMNIST 中手动引入了较高的概念偏移，而其余数据集中的协变量偏移则占据主导地位，这为超参数的选择带来了额外挑战。此外，特征空间的分布对齐有小概率导致反效果：由于目标域采样的随机性，可能存在 $L_t(w) \leq L(w)$。这些因素共同导致了次优的性能表现。

表 5-3 IDM 算法的额外内存与时间开销

数据集	训练时间 (h)			内存需求 (GB)		
	ERM	IDM	额外开销	ERM	IDM	额外开销
ColoredMNIST	0.076	0.088	14.6%	0.138	0.139	0.2%
RotatedMNIST	0.101	0.110	9.3%	0.338	0.342	1.0%
VLCS	0.730	0.744	2.0%	8.189	8.199	0.1%
PACS	0.584	0.593	1.5%	8.189	8.201	0.1%
OfficeHome	0.690	0.710	2.9%	8.191	8.506	3.8%
TerraIncognita	0.829	0.840	1.3%	8.189	8.208	0.2%
DomainNet	2.805	2.947	5.0%	13.406	16.497	23.1%

5.5.3 消融实验

本小节对 IDM 算法的不同组成部分进行消融实验以验证其实际效果，主要针对梯度对齐（GA）、特征对齐（RA）、预热过程（WU）、滑动平均（MA）与 PDM 分布对齐算法开展实验。

表 5-4 ColoredMNIST 数据集上的消融实验

算法	**GA**	**RA**	**WU**	**MA**	**90%**	**80%**	**10%**	**平均准确率**
ERM			-		71.8 ± 0.4	72.9 ± 0.1	28.7 ± 0.5	57.8
IDM	✗	✓	✗	✗	71.9 ± 0.4	72.5 ± 0.0	28.8 ± 0.7	57.7
	✓	✗	✓	✓	73.1 ± 0.2	72.7 ± 0.3	67.4 ± 1.6	71.1
	✓	✓	✗	✓	72.9 ± 0.2	72.7 ± 0.1	60.8 ± 2.1	68.8
	✓	✓	✓	✗	72.0 ± 0.1	71.5 ± 0.3	48.7 ± 7.1	64.0
	✓	✓	✓	✓	**74.2** ± 0.6	**73.5** ± 0.2	**68.3** ± 2.5	**72.0**

表 5-5 OfficeHome 数据集上的消融实验

算法	**GA**	**RA**	**WU**	**MA**	**A**	**C**	**P**	**R**	**平均准确率**
ERM			-		61.7 ± 0.7	53.4 ± 0.3	74.1 ± 0.4	76.2 ± 0.6	66.4
IDM	✗	✓	✗	✗	**64.7** ± 0.5	**54.6** ± 0.3	76.2 ± 0.4	**78.1** ± 0.5	**68.4**
	✓	✗	✓	✓	61.9 ± 0.4	53.0 ± 0.3	75.5 ± 0.2	77.9 ± 0.2	67.1
	✓	✓	✗	✓	62.5 ± 0.1	53.0 ± 0.7	75.0 ± 0.4	77.2 ± 0.7	66.9
	✓	✓	✓	✗	64.2 ± 0.3	53.5 ± 0.6	76.1 ± 0.4	**78.1** ± 0.4	68.0
	✓	✓	✓	✓	64.4 ± 0.3	54.4 ± 0.6	**76.5** ± 0.3	78.0 ± 0.4	68.3

根据上述理论分析，梯度对齐在数据分布包含概念偏移时可促进源域上的泛化。如表 5-4 所示，移除梯度对齐的 IDM 算法（57.7%）性能尚不及朴素的 ERM 方法（57.8%），无法学习不同领域之间的不变性。此外，梯度对齐显著提升了 VLCS（77.4% 到 78.1%）与 PACS（86.8% 到 87.6%）数据集上的性能表现，如表 5-6 与表 5-7 所示。然而，对于

概念偏移并不显著的数据集，例如 OfficeHome，梯度对齐带来的性能提升则并不显著，如表 5-5 所示。

表 5-6 VLCS 数据集上的梯度对齐消融实验

算法	GA	A	C	P	S	平均准确率
ERM	-	**97.6** ± 0.3	**67.9** ± 0.7	70.9 ± 0.2	74.0 ± 0.6	77.6
IDM	✗	97.1 ± 0.7	67.2 ± 0.4	69.9 ± 0.4	75.6 ± 0.8	77.4
IDM	✓	**97.6** ± 0.3	66.9 ± 0.3	**71.8** ± 0.5	**76.0** ± 1.3	**78.1**

表 5-7 PACS 数据集上的梯度对齐消融实验

算法	GA	A	C	P	S	平均准确率
ERM	-	86.5 ± 1.0	81.3 ± 0.6	96.2 ± 0.3	**82.7** ± 1.1	86.7
IDM	✗	87.8 ± 0.6	81.6 ± 0.3	97.4 ± 0.2	80.6 ± 1.3	86.8
IDM	✓	**88.0** ± 0.3	**82.6** ± 0.6	**97.6** ± 0.4	82.3 ± 0.6	**87.6**

特征对齐能够最小化特征空间的协变量偏移，从而促进目标域上的泛化。如表 5-4 与表 5-5 所示，特征对齐在 ColoredMNIST（71.1% 到 72.0%）与 OfficeHome（67.1% 到 68.3%）数据集上均能够显著提高泛化性能。这验证了上述理论结果，即特征对齐补足了梯度对齐，是解决概率领域泛化的充分条件之一。

根据 Arjovsky 等[238,242]的实验设置，仅在训练迭代一定步数之后才加入分布对齐相关损失项，这是由于在训练早期阶段强制模型学习不变性可能阻碍其提取有效信息。通过加入此类预热过程，可使得预测器在训练的开始阶段学习输入—标签间的全部相关性，并在后续更新过程中逐步丢弃不可靠的相关性。如表 5-4 与表 5-5 所示，这一策略有助于提高 ColoredMNIST（68.8% 到 72.0%）与 OfficeHome（66.9% 到 68.3%）数据集上的最终性能。

沿用 Rame 等[242,266]的训练技巧，在分布对齐时引入了滑动平均方法。这一策略有助于在批次大小不足时更为精确地刻画梯度或特征分布。如表 5-4 与表 5-5 所示，滑动平均策略有效提高了 IDM 算法在 ColoredMNIST（64.0% 到 72.0%）与 OfficeHome（68.0% 到 68.3%）数据集上的性能表现。

最后将展示 PDM 分布对齐方法相较于基于低阶矩的分布对齐技术的优越性。具体而言，以下对比了 IGA 算法[240]（对齐梯度期望），Fishr 算法[242]（对齐梯度方差），IGA + Fishr 的组合（同时对齐期望与方差），以及本章的 PDM 算法。最终结果如表 5-8 所示。可见，即便与 IGA + Fishr 组合（70.7%）相比，PDM 依旧显著提高了预测性能（72.0%），这证实了 PDM 方法能够有效应对高维复杂分布的相关讨论。

表 5-8 Colored MNIST 数据集上的 PDM 消融实验

算法	90%	80%	10%	平均准确率
ERM	71.8 ± 0.4	72.9 ± 0.1	28.7 ± 0.5	57.8
IGA	72.6 ± 0.3	72.9 ± 0.2	50.0 ± 1.2	65.2
Fishr	74.1 ± 0.6	73.3 ± 0.1	58.9 ± 3.7	68.8
IGA + Fishr	73.3 ± 0.0	72.6 ± 0.5	66.3 ± 2.9	70.7
IDM	**74.2** ± 0.6	**73.5** ± 0.2	**68.3** ± 2.5	**72.0**

5.6 本章小结

本章针对现有基于数据独立同分布假设的泛化分析框架难以拓展至分布外泛化场景的问题，突破了现有领域泛化理论工作在优化目标刻画方面的局限性，构建了信息论视角下的概率领域泛化分析框架。相较于平均或最坏情形下的分析理论，本框架具备更强的泛化约束性，同时通过松弛相关前提假设，进一步改进了适用范围与可解释性。随后，探索了领域泛化全局总体风险的概率分解，将其拆解为源域与目标域上的泛化问题，并分别建立了信息论视角下的泛化误差上界。基于这些泛化分析结果，发现对齐不同源域中的梯度与特征分布能够有效约束源域与目标域上的泛化误差，进而提出了基于域间分布对齐的领域泛化算法 IDM。进一步地，改进了目前基于低阶矩的分布对齐方法，提出了基于切片与排序的 PDM 分布对齐算法。本章提出的算法在 Colored MNIST 与 DomainBed 基准评估数据集上均取得了优异性能。

本章的研究工作发表于机器学习理论顶级期刊、CCF 推荐 A 类学术期刊 IEEE Transactions on Information Theory，论文题目为“How Does Distribution Matching Help Domain Generalization: An Information-theoretic Analysis”。

6 结论与展望

6.1 研究工作总结

本论文聚焦于信息论视角下随机学习算法的泛化理论研究，针对现有信息论泛化理论工作在可计算性、紧致性以及适用范围等方面的局限性，深入探索研究了传统有监督学习、分布外领域泛化、无监督对比学习等经典学习场景下的泛化理论。本论文的主要研究成果可概括为以下四点：

（1）针对现有基于传统 Shannon 信息度量的信息论泛化误差上界在实际应用中难以量化计算的问题，提出了一种新型信息度量准则，称为核化 Rényi 熵，其能够直接通过有限采样数据点直接近似计算而不受随机变量的维度影响，且同时兼容现有基于传统 Shannon 熵的泛化分析框架。在此之上，成功推广了现有面向随机迭代学习算法的信息论泛化理论，得到了可计算的泛化误差上界估计。通过进一步引入梯度轨迹协方差度量，构建了针对 SGLD 与 SGD 算法的更紧的上界估计结果。进一步地，针对矩阵 Rényi 熵朴素算法计算效率低下的问题，基于矩阵迹估计与多项式近似技术构建了拥有理论最优收敛阶的快速近似算法，为相关泛化上界的可计算性提供切实保障。此研究工作发表于人工智能顶级会议、CCF 推荐 A 类学术会议 International Joint Conference on Artificial Intelligence，论文题目为“Understanding the Generalization Ability of Deep Learning Algorithms: A Kernelized Rényi’s Entropy Perspective”；与机器学习理论顶级期刊、CCF 推荐 A 类学术期刊 IEEE Transactions on Information Theory，论文题目为“Optimal Randomized Approximations for Matrix-based Rényi’s Entropy”。

（2）针对现有基于高维互信息度量的泛化误差上界难以量化计算、估计不紧致的问题，提出了新型低维信息论泛化度量：损失熵，其仅包含一维随机变量，故可通过核密度估计、分箱等方法直接量化计算，从根本上解决了相关泛化上界的不可计算难题。在此之上，成功地将相关理论结果拓展至数据无关泛化场景，显著改进了基于信息瓶颈度量的泛化上界的紧致性与可计算性，同时为最小化误差熵准则的泛化性能提供了全新的理论见解。进一步地，对于数据依赖泛化场景，在留一法与超样本泛化分析框架下改进了现有泛化理论结果，构建了基于损失熵的高概率泛化误差上界。这是首个可计算的高概率信息论泛化上界，且在众多经典学习场景中相较于现有上界估计显著更紧。此研究工作发表于机器学习顶级会议、清华推荐 A 类学术会议 International Conference on Learning Representations，论文题目为“Rethinking Information-theoretic Generalization: Loss Entropy Induced PAC Bounds”。

（3）针对现有面向传统有监督单点学习的信息论泛化上界不再适用于对比学习等多点学习场景的问题，突破了拓展现有理论结果所面临的多项重大挑战，提出了首个

信息论视角下的多点学习泛化理论结果，其适用于任意有界多点损失函数，且能够将包括单点、双点、三点及更高阶情形在内的常见学习范式囊括进统一的泛化分析框架中。首先，针对现有单点学习泛化上界所依赖的损失独立同分布条件在多点学习中不再满足的问题，通过互信息的超可加性进行期望泛化误差的拆分以及自下而上的归约，克服了非独立同分布挑战。其次，针对在超样本框架下推广现有上界所面临的维度爆炸问题，对超样本变量进行独立性拆分，通过低维互信息度量构建了泛化误差上界。此研究工作发表于机器学习顶级会议、CCF 推荐 A 类学术会议 International Conference on Machine Learning，论文题目为“Towards Generalization beyond Pointwise Learning: A Unified Information-theoretic Perspective”。

（4）针对现有基于传统独立同分布假设的泛化理论结果不再适用于领域泛化等分布外泛化场景的问题，突破了现有领域泛化理论工作在优化目标刻画方面的局限性，构建了首个信息论视角下的领域泛化分析框架。首先，通过全局总体风险作为桥梁，将整体领域泛化误差分解为源域与目标域上的泛化误差。进而，分别构建了信息论视角下的源域与目标域泛化误差上界，发现了影响学习算法领域泛化性能的关键互信息度量。随后，成功建立了对齐源域间梯度与特征分布与最小化相关泛化误差上界之间的联系，证明了梯度与特征对齐共同构成领域泛化的一种充分条件。由此，研发了基于域间分布对齐的新型领域泛化算法，并改进了基于低阶矩的分布对齐算法，在多项基准评估数据集上取得了优异性能。此研究工作发表于机器学习理论顶级期刊、CCF 推荐 A 类学术期刊 IEEE Transactions on Information Theory，论文题目为“How Does Distribution Matching Help Domain Generalization: An Information-theoretic Analysis”。

6.2 未来工作展望

虽然本论文在改进现有信息论泛化理论工作的可计算性、紧致性以及适用范围等方面已取得了阶段性成果，但基于信息论的泛化理论分析在可解释性、学习场景以及计算需求等方面仍面临诸多挑战，具备广阔的发展空间。未来工作可从以下方面开展进一步研究：

（1）信息论优化与泛化之间的理论权衡。目前基于信息论的统计机器学习理论主要聚焦于随机学习算法的泛化理论研究，而对神经网络优化过程的探索较少。近期工作[267]表明算法的特定属性或停止策略或将比损失函数或优化方法的选择对于模型性能表现更具影响力。探索深度学习中优化与泛化间的权衡，特别考虑学习率、批次大小或动量等因素对于泛化性能的影响，有望导出更具洞察力的泛化上界。

（2）新型学习场景下的信息论泛化上界。目前基于信息论的泛化分析理论工作主要针对传统有监督学习场景构建，无法自然拓展至半监督等新型学习范式。虽然本论文已在多点学习、领域泛化等拓展场景中探索了相关信息论泛化上界，但其在在线学

习、元学习等新场景下仍具备广阔的发展空间。针对此类新型学习场景对现有理论结果进行拓展与增强，也是信息论泛化研究的重要发展方向。

（3）低估计开销的可计算信息论泛化上界。目前工作中基于信息论的可计算泛化上界多基于 Steinke 等[32]所提出的超样本泛化分析框架。此类上界虽然由于其低维性质能够直接量化计算，但需要额外的验证数据以构成超样本数据集。此外，由于相关互信息度量的计算依赖于概率密度估计，通常需要进行多次独立的模型训练过程以收集足够的数据样本，带来了巨大的计算开销。未来研究可发展低开销的泛化误差上界。

（4）关键泛化度量启发新型学习算法设计。目前的信息论泛化理论主要目标在于解释与估计相关学习算法的泛化行为，例如 Kawaguchi 等[46]解释了信息瓶颈方法的泛化能力，Wang 等[141]则解释了梯度裁剪对于模型泛化的影响。第 5 章展示了信息论泛化上界如何启发基于分布对齐的领域泛化算法设计。此类思想能否拓展至其余关键泛化度量，例如第 2 章中的梯度轨迹协方差，第 3 章中的损失熵，能否通过最小化此类度量启发新型学习算法设计，是一个值得继续探究的问题。

致　谢

首先，我要向我的导师李辰教授和龚铁梁副教授表达我最深切的感谢。感谢李老师在我整个博士研究期间给予的耐心指导、宝贵建议和不懈支持。李老师在学术上的严谨态度和高标准的要求，激励我不断追求卓越，精益求精。龚老师在我研究方向选择以及学术发展上提供了重要的指导，他的智慧和远见帮助我在研究过程中克服了一个又一个难题，使我的研究工作得以顺利推进，并取得了丰硕的成果。

其次，我要感谢我的评审委员会成员以及其他在研究过程中给予我帮助的教授和学者们。你们的反馈和建议不仅提高了我论文的质量，也拓宽了我的学术视野。特别感谢陈洪教授和余书剑教授，你们在我的博士阶段研究过程中提出的宝贵意见与建议，对本论文工作的改进与完善起到了重要的指导作用。

在这里，我还要感谢我的同学和朋友们，感谢你们在我遇到困难时给予的鼓励和支持，在我取得进展时与我一同分享喜悦。感谢高泽宇、和凯、吴佳伦、娄沛良、时江波、March、李羽霏、廉雨辰、马骁勇、洪邦洋、袁义和、毛安钰、刘佳帅、王翔宇、张泽扬等实验室的同学们，也感谢张博航、瞿建、陈志扬等长久以来的好伙伴，我们一起共度了无数个研究和讨论的日夜，正是这些共同努力的时光，使我的博士生活充实而有意义。感谢我的家人，感谢我的父母，你们无条件的爱和支持是我坚持不懈、勇往直前的动力源泉。你们的理解和鼓励使我在面对学术挑战时能够始终保持积极乐观。

在这段博士研究的旅程中，我经历了许多挑战和困难，但也收获了许多宝贵的经验和知识。这些经历不仅让我在学术上取得了进步，也让我在个人成长中得到了升华。我深知，今天取得的这些成就，离不开每一位在我学术道路上给予帮助和支持的人。你们的关心和支持让我深感幸运，也让我充满感激之情。

再次衷心感谢你们，正是有了你们的帮助和支持，我才能完成这篇博士论文，并在学术道路上迈出坚实的一步。未来，我将继续努力，不辜负你们的期望，为科学研究和社会发展贡献自己的绵薄之力。

参考文献

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

[2] SHALEV-SHWARTZ S, BEN-DAVID S. Understanding machine learning: From theory to algorithms[M]. Cambridge university press, 2014.

[3] VAPNIK V. The nature of statistical learning theory[M]. Springer science & business media, 2013.

[4] VAPNIK V N, CHERVONENKIS A Y. On the uniform convergence of relative frequencies of events to their probabilities[G]//Measures of complexity: festschrift for alexey chervonenkis. Springer, 2015: 11-30.

[5] BLUMER A, EHRENFEUCHT A, HAUSSLER D, et al. Learnability and the Vapnik-Chervonenkis dimension[J]. Journal of the ACM (JACM), 1989, 36(4): 929-965.

[6] BARTLETT P L, BOUSQUET O, MENDELSON S. Local Rademacher Complexities[J]. The Annals of Statistics, 2005, 33(4): 1497-1537.

[7] MOHRI M, ROSTAMIZADEH A. Rademacher complexity bounds for non-iid processes[J]. Advances in Neural Information Processing Systems, 2008, 21.

[8] ALLEN-ZHU Z, LI Y, SONG Z. A convergence theory for deep learning via over-parameterization [C]//International conference on machine learning. 2019: 242-252.

[9] BELKIN M, HSU D, XU J. Two models of double descent for weak features[J]. SIAM Journal on Mathematics of Data Science, 2020, 2(4): 1167-1180.

[10] BOUSQUET O, ELISSEEFF A. Stability and generalization[J]. The Journal of Machine Learning Research, 2002, 2: 499-526.

[11] FELDMAN V, VONDRAK J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate[C]//Conference on Learning Theory. 2019: 1270-1279.

[12] LIU T, LUGOSI G, NEU G, et al. Algorithmic stability and hypothesis complexity[C]// International Conference on Machine Learning. 2017: 2159-2167.

[13] LEI Y, YING Y. Fine-grained analysis of stability and generalization for stochastic gradient descent [C]//International Conference on Machine Learning. 2020: 5809-5819.

[14] SHALEV-SHWARTZ S, SHAMIR O, SREBRO N, et al. Learnability, stability and uniform convergence[J]. The Journal of Machine Learning Research, 2010, 11: 2635-2670.

[15] JEON H J, VAN ROY B. An information-theoretic framework for deep learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 3279-3291.

[16] BLUMER A, EHRENFEUCHT A, HAUSSLER D, et al. Occam's razor[J]. Information processing letters, 1987, 24(6): 377-380.

[17] EDGEWORTH F Y. On the probable errors of frequency-constants[J]. Journal of the Royal Statistical Society, 1908, 71(2): 381-397.

[18] FISHER R A. On the mathematical foundations of theoretical statistics[J]. Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical

character, 1922, 222(594-604): 309-368.

[19] SHANNON C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[20] KOLMOGOROV A N. On tables of random numbers[J]. Sankhyā: The Indian Journal of Statistics, Series A, 1963: 369-376.

[21] YANG Y, BARRON A. Information-theoretic determination of minimax rates of convergence[J]. Annals of Statistics, 1999: 1564-1599.

[22] LEUNG G, BARRON A R. Information theory and mixing least-squares regressions[J]. IEEE Transactions on information theory, 2006, 52(8): 3396-3410.

[23] AKAIKE H. A new look at the statistical model identification[J]. IEEE transactions on automatic control, 1974, 19(6): 716-723.

[24] SCHWARZ G. Estimating the dimension of a model[J]. The annals of statistics, 1978: 461-464.

[25] RISSANEN J. Modeling by shortest data description[J]. Automatica, 1978, 14(5): 465-471.

[26] ZHANG T. Information-theoretic upper and lower bounds for statistical estimation[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1307-1321.

[27] RUSSO D, ZOU J. Controlling bias in adaptive data analysis using information theory[C]// Artificial Intelligence and Statistics. 2016: 1232-1240.

[28] XU A, RAGINSKY M. Information-theoretic analysis of generalization capability of learning algorithms[J]. Advances in neural information processing systems, 2017, 30.

[29] MCALLESTER D A. Some PAC-Bayesian theorems[C]//Proceedings of the eleventh annual conference on Computational learning theory. 1998: 230-234.

[30] SHAWE-TAYLOR J, WILLIAMSON R C. A PAC analysis of a Bayesian estimator[C]// Proceedings of the tenth annual conference on Computational learning theory. 1997: 2-9.

[31] CATONI O. Pac-Bayesian supervised classification: The thermodynamics of statistical learning[J]. stat, 2007, 1050: 3.

[32] STEINKE T, ZAKYNTHINOU L. Reasoning about generalization via conditional mutual information[C]//Conference on Learning Theory. 2020: 3437-3452.

[33] NEGREA J, HAGHIFAM M, DZIUGAITE G K, et al. Information-theoretic generalization bounds for SGLD via data-dependent estimates[J]. Advances in Neural Information Processing Systems, 2019, 32.

[34] NEU G, DZIUGAITE G K, HAGHIFAM M, et al. Information-theoretic generalization bounds for stochastic gradient descent[C]//Conference on Learning Theory. 2021: 3526-3545.

[35] DONSKER M D, VARADHAN S S. Asymptotic evaluation of certain Markov process expectations for large time, I[J]. Communications on pure and applied mathematics, 1975, 28(1): 1-47.

[36] RAGINSKY M, RAKHLIN A, TSAO M, et al. Information-theoretic analysis of stability and bias of learning algorithms[C]//2016 IEEE Information Theory Workshop (ITW). 2016: 26-30.

[37] LANGFORD J, SEEGER M. Bounds for averaging classifiers[M]. School of Computer Science, Carnegie Mellon University, 2001.

[38] MCALLESTER D A. PAC-Bayesian stochastic model selection[J]. Machine Learning, 2003, 51(1):

5-21.

[39] AUDIBERT J Y. A better variance control for PAC-Bayesian classification[J]. Preprint, 2004, 905.

[40] TISHBY N, PEREIRA F C, BIALEK W. The information bottleneck method[J]. arXiv preprint physics/0004057, 2000.

[41] SHWARTZ-ZIV R, TISHBY N. Opening the black box of deep neural networks via information[J]. arXiv preprint arXiv:1703.00810, 2017.

[42] ACHILLE A, SOATTO S. Emergence of invariance and disentanglement in deep representations [J]. Journal of Machine Learning Research, 2018, 19(50): 1-34.

[43] SAXE A M, BANSAL Y, DAPELLO J, et al. On the information bottleneck theory of deep learning [J]. Journal of Statistical Mechanics: Theory and Experiment, 2019, 2019(12): 124020.

[44] GOLDFELD Z, POLYANSKIY Y. The information bottleneck problem and its applications in machine learning[J]. IEEE Journal on Selected Areas in Information Theory, 2020, 1(1): 19-38.

[45] GEIGER B C. On information plane analyses of neural network classifiers—A review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(12): 7039-7051.

[46] KAWAGUCHI K, DENG Z, JI X, et al. How does information bottleneck help deep learning?[C]//International Conference on Machine Learning. 2023: 16049-16096.

[47] BÉGIN L, GERMAIN P, LAVIOLETTE F, et al. PAC-Bayesian bounds based on the Rényi divergence[C]//Artificial Intelligence and Statistics. 2016: 435-444.

[48] ALQUIER P, GUEDJ B. Simpler PAC-Bayesian bounds for hostile data[J]. Machine Learning, 2018, 107(5): 887-902.

[49] RIVASPLATA O, KUZBORSKIJ I, SZEPESVÁRI C, et al. PAC-Bayes analysis beyond the usual bounds[J]. Advances in Neural Information Processing Systems, 2020, 33: 16833-16845.

[50] RODRÍGUEZ-GÁLVEZ B, BASSI G, THOBABEN R, et al. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm[C]//2020 IEEE Information Theory Workshop (ITW). 2021: 1-5.

[51] HELLSTRÖM F, DURISI G. A new family of generalization bounds using samplewise evaluated CMI[J]. Advances in Neural Information Processing Systems, 2022, 35: 10108-10121.

[52] MCALLESTER D. A PAC-Bayesian tutorial with a dropout bound[J]. arXiv preprint arXiv:1307.2118, 2013.

[53] JIAO J, HAN Y, WEISSMAN T. Dependence measures bounding the exploration bias for general measurements[C]//2017 IEEE International Symposium on Information Theory (ISIT). 2017: 1475-1479.

[54] BU Y, ZOU S, VEERAVALLI V V. Tightening mutual information-based bounds on generalization error[J]. IEEE Journal on Selected Areas in Information Theory, 2020, 1(1): 121-130.

[55] HAGHIFAM M, NEGREA J, KHISTI A, et al. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms[J]. Advances in Neural Information Processing Systems, 2020, 33: 9925-9935.

[56] HELLSTRÖM F, DURISI G. Data-dependent PAC-Bayesian bounds in the random-subset setting with applications to neural networks[C]//Workshop on Inf.-Theoretic Methods Rigorous, Respon-

sible, and Reliable Mach. Learn.(ITR3), Virtual conference. 2021.

[57] ZHOU R, TIAN C, LIU T. Individually conditional individual mutual information bound on generalization error[J]. IEEE Transactions on Information Theory, 2022, 68(5): 3304-3316.

[58] HARUTYUNYAN H, RAGINSKY M, VER STEEG G, et al. Information-theoretic generalization bounds for black-box learning algorithms[J]. Advances in Neural Information Processing Systems, 2021, 34: 24670-24682.

[59] HARUTYUNYAN H, VER STEEG G, GALSTYAN A. Formal limitations of sample-wise information-theoretic generalization bounds[C]//2022 IEEE Information Theory Workshop (ITW). 2022: 440-445.

[60] AMINIAN G, ABROSHAN M, KHALILI M M, et al. An information-theoretical approach to semi-supervised learning under covariate-shift[C]//International Conference on Artificial Intelligence and Statistics. 2022: 7433-7449.

[61] WINTENBERGER O. Weak transport inequalities and applications to exponential and oracle inequalities[J]. 2015.

[62] LOPEZ A T, JOG V. Generalization error bounds using Wasserstein distances[C]//2018 IEEE Information Theory Workshop (ITW). 2018: 1-5.

[63] WANG H, DIAZ M, SANTOS FILHO J C S, et al. An information-theoretic view of generalization via Wasserstein distance[C]//2019 IEEE International Symposium on Information Theory (ISIT). 2019: 577-581.

[64] RODRÍGUEZ GÁLVEZ B, BASSI G, THOBABEN R, et al. Tighter expected generalization error bounds via Wasserstein distance[J]. Advances in Neural Information Processing Systems, 2021, 34: 19109-19121.

[65] CLERICO E, SHIDANI A, DELIGIANNIDIS G, et al. Chained generalisation bounds[C]// Conference on Learning Theory. 2022: 4212-4257.

[66] ALABDULMOHSIN I. Towards a unified theory of learning and information[J]. Entropy, 2020, 22(4): 438.

[67] HAFEZ-KOLAHI H, GOLGOONI Z, KASAEI S, et al. Conditioning and processing: Techniques to improve information-theoretic generalization bounds[J]. Advances in Neural Information Processing Systems, 2020, 33: 16457-16467.

[68] AMINIAN G, TONI L, RODRIGUES M R. Jensen-Shannon information based characterization of the generalization error of learning algorithms[C]//2020 IEEE Information Theory Workshop (ITW). 2021: 1-5.

[69] MODAK E, ASNANI H, PRABHAKARAN V M. Rényi divergence based bounds on generalization error[C]//2021 IEEE Information Theory Workshop (ITW). 2021: 1-6.

[70] AMINIAN G, TONI L, RODRIGUES M R. Information-theoretic bounds on the moments of the generalization error of learning algorithms[C]//2021 IEEE International Symposium on Information Theory (ISIT). 2021: 682-687.

[71] RAGINSKY M, RAKHLIN A, XU A. 10 Information-Theoretic Stability and Generalization[J]. Information-Theoretic Methods in Data Science, 2021: 302.

[72] SEFIDGARAN M, GOHARI A, RICHARD G, et al. Rate-distortion theoretic generalization bounds for stochastic learning algorithms[C]//Conference on Learning Theory. 2022: 4416-4463.

[73] ESPOSITO A R, GASTPAR M. From generalisation error to transportation-cost inequalities and back[C]//2022 IEEE International Symposium on Information Theory (ISIT). 2022: 294-299.

[74] WONGSO S, GHOSH R, MOTANI M. Understanding deep neural networks using sliced mutual information[C]//2022 IEEE International Symposium on Information Theory (ISIT). 2022: 133-138.

[75] WONGSO S, GHOSH R, MOTANI M. Using sliced mutual information to study memorization and generalization in deep neural networks[C]//International Conference on Artificial Intelligence and Statistics. 2023: 11608-11629.

[76] CHU Y, RAGINSKY M. A unified framework for information-theoretic generalization bounds[J]. Advances in Neural Information Processing Systems, 2024, 36.

[77] HAFEZ-KOLAHI H, MONIRI B, KASAEI S, et al. Rate-distortion analysis of minimum excess risk in Bayesian learning[C]//International Conference on Machine Learning. 2021: 3998-4007.

[78] XU A, RAGINSKY M. Minimum excess risk in Bayesian learning[J]. IEEE Transactions on Information Theory, 2022, 68(12): 7935-7955.

[79] HAFEZ-KOLAHI H, MONIRI B, KASAEI S. Information-theoretic analysis of minimax excess risk[J]. IEEE Transactions on Information Theory, 2023.

[80] DOGAN M B, GASTPAR M. Lower bounds on the expected excess risk using mutual information [C]//2021 IEEE Information Theory Workshop (ITW). 2021: 1-6.

[81] KOOLEN W M, GRÜNWALD P, VAN ERVEN T. Combining adversarial guarantees and stochastic fast rates in online learning[J]. Advances in Neural Information Processing Systems, 2016, 29.

[82] MHAMMEDI Z, GRÜNWALD P, GUEDJ B. PAC-Bayes un-expected Bernstein inequality[J]. Advances in Neural Information Processing Systems, 2019, 32.

[83] GRÜNWALD P D, MEHTA N A. Fast rates for general unbounded loss functions: from ERM to generalized Bayes[J]. Journal of Machine Learning Research, 2020, 21(56): 1-80.

[84] GRÜNWALD P D, PÉREZ-ORTIZ M F, MHAMMEDI Z. Exponential stochastic inequality[J]. arXiv preprint arXiv:2304.14217, 2023.

[85] GERMAIN P, LACASSE A, LAVIOLETTE F, et al. PAC-Bayesian learning of linear classifiers[C]//Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 353-360.

[86] BÉGIN L, GERMAIN P, LAVIOLETTE F, et al. PAC-Bayesian theory for transductive learning [C]//Artificial Intelligence and Statistics. 2014: 105-113.

[87] HELLSTRÖM F, DURISI G. Generalization bounds via information density and conditional information density[J]. IEEE Journal on Selected Areas in Information Theory, 2020, 1(3): 824-839.

[88] MAURER A. A note on the PAC Bayesian theorem[J]. arXiv preprint cs/0411099, 2004.

[89] FOONG A, BRUINSMA W, BURT D, et al. How tight can PAC-Bayes be in the small data regime? [J]. Advances in Neural Information Processing Systems, 2021, 34: 4093-4105.

[90] HELLSTRÖM F, GUEDJ B. Comparing comparators in generalization bounds[C]//International Conference on Artificial Intelligence and Statistics. 2024: 73-81.

[91] JANG K, JUN K S, KUZBORSKIJ I, et al. Tighter PAC-Bayes bounds through coin-betting[C]// The Thirty Sixth Annual Conference on Learning Theory. 2023: 2240-2264.

[92] OHNISHI Y, HONORIO J. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation[C]//International conference on artificial intelligence and statistics. 2021: 1711-1719.

[93] AMBROLADZE A, PARRADO-HERNÁNDEZ E, SHAWE-TAYLOR J. Tighter PAC-Bayes bounds[J]. Advances in neural information processing systems, 2006, 19.

[94] SHAWE-TAYLOR J, PARRADO-HERNÁNDEZ E, AMBROLADZE A. Data dependent priors in PAC-Bayes bounds[C]//Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers. 2010: 231-240.

[95] DZIUGAITE G K, ROY D M. Data-dependent PAC-Bayes priors via differential privacy[J]. Advances in neural information processing systems, 2018, 31.

[96] RIVASPLATA O, PARRADO-HERNÁNDEZ E, SHAWE-TAYLOR J S, et al. PAC-Bayes bounds for stable algorithms with instance-dependent priors[J]. Advances in Neural Information Processing Systems, 2018, 31.

[97] DZIUGAITE G K, HSU K, GHARBIEH W, et al. On the role of data in PAC-Bayes bounds[C]// International Conference on Artificial Intelligence and Statistics. 2021: 604-612.

[98] SEEGER M. PAC-Bayesian generalisation error bounds for Gaussian process classification[J]. Journal of machine learning research, 2002, 3(Oct): 233-269.

[99] LEVER G, LAVIOLETTE F, SHAWE-TAYLOR J. Distribution-dependent PAC-Bayes priors[C] //International Conference on Algorithmic Learning Theory. 2010: 119-133.

[100] LEVER G, LAVIOLETTE F, SHAWE-TAYLOR J. Tighter PAC-Bayes bounds through distribution-dependent priors[J]. Theoretical Computer Science, 2013, 473: 4-28.

[101] HELLSTRÖM F, DURISI G. Fast-rate loss bounds via conditional information measures with applications to neural networks[C]//2021 IEEE International Symposium on Information Theory (ISIT). 2021: 952-957.

[102] ESPOSITO A R, GASTPAR M, ISSA I. Generalization error bounds via Rényi-, f-divergences and maximal leakage[J]. IEEE Transactions on Information Theory, 2021, 67(8): 4986-5004.

[103] BASSILY R, NISSIM K, SMITH A, et al. Algorithmic stability for adaptive data analysis[C]// Proceedings of the forty-eighth annual ACM symposium on Theory of Computing. 2016: 1046-1059.

[104] GERMAIN P, BACH F, LACOSTE A, et al. PAC-Bayesian theory meets Bayesian inference[J]. Advances in Neural Information Processing Systems, 2016, 29.

[105] CATONI O. Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001[M]. Springer Science & Business Media, 2004.

[106] ALQUIER P. Transductive and inductive adaptative inference for regression and density estimation [D]. ENSAE ParisTech, 2006.

[107] ALQUIER P. PAC-Bayesian bounds for randomized empirical risk minimizers[J]. Mathematical

Methods of Statistics, 2008, 17: 279-304.

[108] CATONI O, GIULINI I. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector[J]. arXiv preprint arXiv:1802.04308, 2018.

[109] HOLLAND M. PAC-Bayes under potentially heavy tails[J]. Advances in Neural Information Processing Systems, 2019, 32.

[110] BIGGS F, GUEDJ B. Tighter pac-bayes generalisation bounds by leveraging example difficulty[C]//International Conference on Artificial Intelligence and Statistics. 2023: 8165-8182.

[111] HERBRICH R, GRAEPEL T. A PAC-Bayesian margin bound for linear classifiers[J]. IEEE Transactions on Information Theory, 2002, 48(12): 3140-3150.

[112] LANGFORD J, SHAWE-TAYLOR J. PAC-Bayes & margins[J]. Advances in neural information processing systems, 2002, 15.

[113] BIGGS F, GUEDJ B. On margins and derandomisation in PAC-Bayes[C]//International Conference on Artificial Intelligence and Statistics. 2022: 3709-3731.

[114] AUDIBERT J Y, BOUSQUERT O. Combining PAC-Bayesian and generic chaining bounds[J]. Journal of Machine Learning Research, 2007, 8(4).

[115] ASADI A R, ABBE E. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks[J]. Journal of Machine Learning Research, 2020, 21(139): 1-32.

[116] YANG J, SUN S, ROY D M. Fast-rate PAC-Bayes generalization bounds via shifted Rademacher processes[J]. Advances in Neural Information Processing Systems, 2019, 32.

[117] SAUNSHI N, PLEVRAKIS O, ARORA S, et al. A theoretical analysis of contrastive unsupervised representation learning[C]//International Conference on Machine Learning. 2019: 5628-5637.

[118] NOZAWA K, GERMAIN P, GUEDJ B. PAC-Bayesian contrastive unsupervised representation learning[C]//Conference on Uncertainty in Artificial Intelligence. 2020: 21-30.

[119] MHAMMEDI Z, GUEDJ B, WILLIAMSON R C. Pac-bayesian bound for the conditional value at risk[J]. Advances in Neural Information Processing Systems, 2020, 33: 17919-17930.

[120] CHÉRIEF-ABDELLATIF B E, SHI Y, DOUCET A, et al. On PAC-Bayesian reconstruction guarantees for VAEs[C]//International conference on artificial intelligence and statistics. 2022: 3066-3079.

[121] MBACKE S D, CLERC F, GERMAIN P. Statistical guarantees for variational autoencoders using pac-bayesian theory[J]. Advances in Neural Information Processing Systems, 2024, 36.

[122] MBACKE S D, CLERC F, GERMAIN P. PAC-Bayesian generalization bounds for adversarial generative models[C]//International Conference on Machine Learning. 2023: 24271-24290.

[123] HADDOUCHE M, GUEDJ B, RIVASPLATA O, et al. PAC-Bayes unleashed: Generalisation bounds with unbounded losses[J]. Entropy, 2021, 23(10): 1330.

[124] HADDOUCHE M, GUEDJ B. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales[J]. Transactions on Machine Learning Research, 2023.

[125] HADDOUCHE M, GUEDJ B. Wasserstein PAC-Bayes learning: Exploiting optimisation guarantees to explain generalisation[J]. arXiv preprint arXiv:2304.07048, 2023.

[126] VIALLARD P, HADDOUCHE M, SIMSEKLI U, et al. Learning via Wasserstein-based high prob-

ability generalisation bounds[J]. Advances in Neural Information Processing Systems, 2024, 36.

[127] GRUNWALD P, STEINKE T, ZAKYNTHINOU L. PAC-Bayes, MAC-Bayes and conditional mutual information: Fast rate bounds that handle general VC classes[C]//Conference on Learning Theory. 2021: 2217-2247.

[128] HAGHIFAM M, MORAN S, ROY D M, et al. Understanding generalization via leave-one-out conditional mutual information[C]//2022 IEEE International Symposium on Information Theory (ISIT). 2022: 2487-2492.

[129] RAMMAL M R, ACHILLE A, GOLATKAR A, et al. On leave-one-out conditional mutual information for generalization[J]. Advances in Neural Information Processing Systems, 2022, 35: 10179-10190.

[130] WANG Z, MAO Y. Tighter information-theoretic generalization bounds from supersamples[C]// International Conference on Machine Learning. 2023: 36111-36137.

[131] WANG Z, MAO Y. Sample-conditioned hypothesis stability sharpens information-theoretic generalization bounds[J]. Advances in Neural Information Processing Systems, 2024, 36.

[132] SACHS S, van ERVEN T, HODGKINSON L, et al. Generalization guarantees via algorithm-dependent rademacher complexity[C]//The Thirty Sixth Annual Conference on Learning Theory. 2023: 4863-4880.

[133] SEFIDGARAN M, ZAIDI A, KRASNOWSKI P. Minimum description length and generalization guarantees for representation learning[J]. Advances in Neural Information Processing Systems, 2024, 36.

[134] PENSIA A, JOG V, LOH P L. Generalization error bounds for noisy, iterative algorithms[C]// 2018 IEEE International Symposium on Information Theory (ISIT). 2018: 546-550.

[135] MOU W, WANG L, ZHAI X, et al. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints[C]//Conference on Learning Theory. 2018: 605-638.

[136] LI J, LUO X, QIAO M. On generalization error bounds of noisy gradient methods for non-convex learning[C]//International Conference on Learning Representations. 2019.

[137] WANG B, ZHANG H, ZHANG J, et al. Optimizing information-theoretical generalization bound via anisotropic noise of SGLD[J]. Advances in Neural Information Processing Systems, 2021, 34: 26080-26090.

[138] WANG H, HUANG Y, GAO R, et al. Analyzing the generalization capability of SGLD using properties of Gaussian channels[J]. Advances in Neural Information Processing Systems, 2021, 34: 24222-24234.

[139] FUTAMI F, FUJISAWA M. Time-independent information-theoretic generalization bounds for SGLD[J]. Advances in Neural Information Processing Systems, 2024, 36.

[140] ISSA I, ESPOSITO A R, GASTPAR M. Generalization error bounds for noisy, iterative algorithms via maximal leakage[J]. arXiv preprint arXiv:2302.14518, 2023.

[141] WANG Z, MAO Y. On the generalization of models trained with SGD: information-theoretic bounds and implications[C]//International Conference on Learning Representations. 2021.

[142] HAGHIFAM M, RODRÍGUEZ-GÁLVEZ B, THOBABEN R, et al. Limitations of information-

theoretic generalization bounds for gradient descent methods in stochastic convex optimization[C]//International Conference on Algorithmic Learning Theory. 2023: 663-706.

[143] NEYSHABUR B, TOMIOKA R, SREBRO N. Norm-based capacity control in neural networks[C]//Conference on learning theory. 2015: 1376-1401.

[144] BARTLETT P L, FOSTER D J, TELGARSKY M J. Spectrally-normalized margin bounds for neural networks[J]. Advances in neural information processing systems, 2017, 30.

[145] NEYSHABUR B, BHOJANAPALLI S, SREBRO N. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks[C]//International Conference on Learning Representations. 2018.

[146] FORET P, KLEINER A, MOBAHI H, et al. Sharpness-aware minimization for efficiently improving generalization[C]//International Conference on Learning Representations. 2020.

[147] TSUZUKU Y, SATO I, SUGIYAMA M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis[C]//International Conference on Machine Learning. 2020: 9636-9647.

[148] BANERJEE A, CHEN T, LI X, et al. Stability based generalization bounds for exponential family langevin dynamics[C]//International Conference on Machine Learning. 2022: 1412-1449.

[149] PITAS K. Dissecting non-vacuous generalization bounds based on the mean-field approximation [C]//International Conference on Machine Learning. 2020: 7739-7749.

[150] DZIUGAITE G K, ROY D. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors[C]//International Conference on Machine Learning. 2018: 1377-1386.

[151] LEE J, BAHRI Y, NOVAK R, et al. Deep neural networks as Gaussian processes[C]//International Conference on Learning Representations. 2018.

[152] VALLE-PEREZ G, CAMARGO C Q, LOUIS A A. Deep learning generalizes because the parameter-function map is biased towards simple functions[C]//International Conference on Learning Representations. 2018.

[153] BERNSTEIN J, YUE Y. Computing the information content of trained neural networks[J]. arXiv preprint arXiv:2103.01045, 2021.

[154] JACOT A, GABRIEL F, HONGLER C. Neural tangent kernel: Convergence and generalization in neural networks[J]. Advances in neural information processing systems, 2018, 31.

[155] SHWARTZ-ZIV R, ALEMI A A. Information in infinite ensembles of infinitely-wide neural networks[C]//Symposium on Advances in Approximate Bayesian Inference. 2020: 1-17.

[156] CLERICO E, DELIGIANNIDIS G, DOUCET A. Wide stochastic networks: Gaussian limit and PAC-Bayesian training[C]//International Conference on Algorithmic Learning Theory. 2023: 447-470.

[157] HUANG W, LIU C, CHEN Y, et al. Analyzing deep pac-bayesian learning with neural tangent kernel: Convergence, analytic generalization bound, and efficient hyperparameter selection[J]. Transactions on Machine Learning Research, 2023.

[158] CLERICO E, GUEDJ B. A note on regularised NTK dynamics with an application to PAC-Bayesian

training[J]. arXiv preprint arXiv:2312.13259, 2023.

[159] WANG Z, HUANG S L, KURUOGLU E E, et al. PAC-Bayes information bottleneck[C]// International Conference on Learning Representations. 2021.

[160] VIALLARD P, EMONET R, GERMAIN P, et al. Interpreting neural networks as majority votes through the PAC-Bayesian theory[C]//Workshop on Machine Learning with guarantees@ NeurIPS 2019. 2019.

[161] RIVASPLATA O, TANKASALI V M, SZEPESVARI C. PAC-Bayes with backprop[J]. arXiv preprint arXiv:1908.07380, 2019.

[162] LETARTE G, GERMAIN P, GUEDJ B, et al. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32.

[163] BIGGS F, GUEDJ B. Differentiable pac–bayes objectives with partially aggregated neural networks [J]. Entropy, 2021, 23(10): 1280.

[164] BIGGS F, GUEDJ B. Non-vacuous generalisation bounds for shallow neural networks[C]// International Conference on Machine Learning. 2022: 1963-1981.

[165] ZANTEDESCHI V, VIALLARD P, MORVANT E, et al. Learning stochastic majority votes by minimizing a PAC-Bayes generalization bound[J]. Advances in Neural Information Processing Systems, 2021, 34: 455-467.

[166] NAGARAJAN V, KOLTER Z. Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience[C]//International Conference on Learning Representations. 2018.

[167] TINSI L, DALALYAN A. Risk bounds for aggregated shallow neural networks using Gaussian priors[C]//Conference on Learning Theory. 2022: 227-253.

[168] CLERICO E, DELIGIANNIDIS G, DOUCET A. Conditionally gaussian pac-bayes[C]// International Conference on Artificial Intelligence and Statistics. 2022: 2311-2329.

[169] JIN G, YI X, YANG P, et al. Weight expansion: A new perspective on dropout and generalization [J]. Transactions on Machine Learning Research, 2022.

[170] LIAO R, URTASUN R, ZEMEL R. A PAC-Bayesian approach to generalization bounds for graph neural networks[C]//International Conference on Learning Representations. 2020.

[171] VIALLARD P, VIDOT E G, HABRARD A, et al. A pac-bayes analysis of adversarial robustness [J]. Advances in Neural Information Processing Systems, 2021, 34: 14421-14433.

[172] XIAO J, SUN R, LUO Z Q. PAC-bayesian spectrally-normalized bounds for adversarially robust generalization[J]. Advances in Neural Information Processing Systems, 2023, 36: 36305-36323.

[173] ZHANG C, BENGIO S, HARDT M, et al. Understanding deep learning (still) requires rethinking generalization[J]. Communications of the ACM, 2021, 64(3): 107-115.

[174] HARDT M, RECHT B, SINGER Y. Train faster, generalize better: Stability of stochastic gradient descent[C]//International conference on machine learning. 2016: 1225-1234.

[175] LI S, LIU Y. High probability guarantees for nonconvex stochastic gradient descent with heavy tails[C]//International Conference on Machine Learning. 2022: 12931-12963.

[176] BOTTOU L, CURTIS F E, NOCEDAL J. Optimization methods for large-scale machine learning

[J]. Siam Review, 2018, 60(2): 223-311.

[177] BASSILY R, FELDMAN V, GUZMÁN C, et al. Stability of stochastic gradient descent on nonsmooth convex losses[J]. Advances in Neural Information Processing Systems, 2020, 33: 4381-4391.

[178] LEI Y, YANG Z, YANG T, et al. Stability and generalization of stochastic gradient methods for minimax problems[C]//International Conference on Machine Learning. 2021: 6175-6186.

[179] YANG Z, LEI Y, WANG P, et al. Simple stochastic and online gradient descent algorithms for pairwise learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 20160-20171.

[180] YANG Z, LEI Y, LYU S, et al. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss[C]//International Conference on Artificial Intelligence and Statistics. 2021: 2026-2034.

[181] ARORA S, GE R, NEYSHABUR B, et al. Stronger generalization bounds for deep nets via a compression approach[C]//International Conference on Machine Learning. 2018: 254-263.

[182] ZHOU W, VEITCH V, AUSTERN M, et al. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach[C]//Advances in Neural Information Processing Systems. 2018.

[183] LEI Y, HU T, TANG K. Generalization performance of multi-pass stochastic gradient descent with convex loss functions[J]. Journal of Machine Learning Research, 2021, 22: 1-41.

[184] RÉNYI A. On measures of entropy and information[C]//Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. 1961: 547-561.

[185] GIRALDO L G S, RAO M, PRINCIPE J C. Measures of entropy from data using infinitely divisible kernels[J]. IEEE Transactions on Information Theory, 2014, 61(1): 535-548.

[186] YU S, GIRALDO L G S, JENSSEN R, et al. Multivariate extension of matrix-based Rényi's α-order entropy functional[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(11): 2960-2966.

[187] AVRON H, TOLEDO S. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix[J]. Journal of the ACM (JACM), 2011, 58(2): 1-34.

[188] MEYER R A, MUSCO C, MUSCO C, et al. Hutch++: Optimal stochastic trace estimation[J]. Symposium on Simplicity in Algorithms (SOSA), 2021: 142-155. arXiv: 2010.09649.

[189] GURBUZBALABAN M, SIMSEKLI U, ZHU L. The heavy-tail phenomenon in SGD[C]// International Conference on Machine Learning. 2021: 3964-3975.

[190] CAMUTO A, WANG X, ZHU L, et al. Asymmetric heavy tails and implicit bias in gaussian noise injections[C]//International Conference on Machine Learning. 2021: 1249-1260.

[191] MAHONEY M, MARTIN C. Traditional and heavy tailed self regularization in neural network models[C]//International Conference on Machine Learning. 2019: 4284-4293.

[192] MARTIN C H, MAHONEY M W. Implicit self-regularization in deep neural networks: evidence from random matrix theory and implications for learning.[J]. J. Mach. Learn. Res., 2021, 22(165): 1-73.

[193] BASSILY R, MORAN S, NACHUM I, et al. Learners that use little information[C]//Algorithmic

Learning Theory. 2018: 25-55.

[194] RAGINSKY M, RAKHLIN A, XU A. Information-theoretic stability and generalization[J]. Information-Theoretic Methods in Data Science, 2021: 302.

[195] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning[C]//2019 IEEE symposium on security and privacy (SP). 2019: 739-753.

[196] PRINCIPE J C. Information theoretic learning: Renyi's entropy and kernel perspectives[M]. Springer Science & Business Media, 2010.

[197] RAMMAL M R, ACHILLE A, GOLATKAR A, et al. On leave-one-out conditional mutual information for generalization[C]//OH A H, AGARWAL A, BELGRAVE D, et al. Advances in Neural Information Processing Systems. 2022.

[198] ZHIVOTOVSKIY N, HANNEKE S. Localization of VC classes: Beyond local Rademacher complexities[J]. Theoretical Computer Science, 2018, 742: 27-49.

[199] AMJAD R A, GEIGER B C. Learning representations for neural network-based classification using the information bottleneck principle[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(9): 2225-2239.

[200] LECUN Y, CORTES C. MNIST handwritten digit database[EB/OL]. 2010. http://yann.lecun.com/exdb/mnist/.

[201] KRIZHEVSKY A, HINTON G, et al. Learning multiple layers of features from tiny images[Z]. 2009.

[202] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[203] KINGMA D P, SALIMANS T, WELLING M. Variational dropout and the local reparameterization trick[J]. Advances in neural information processing systems, 2015, 28.

[204] MANDT S, HOFFMAN M D, BLEI D M. Stochastic gradient descent as approximate bayesian inference[J]. Journal of Machine Learning Research, 2017, 18: 1-35.

[205] MADDOX W J, IZMAILOV P, GARIPOV T, et al. A simple baseline for bayesian uncertainty in deep learning[J]. Advances in neural information processing systems, 2019, 32.

[206] GALLOWAY A, GOLUBEVA A, SALEM M, et al. Bounding generalization error with input compression: An empirical study with infinite-width networks[J]. Transactions on Machine Learning Research, 2023.

[207] CHEN T, KORNBLITH S, NOROUZI M, et al. Simclr: A simple framework for contrastive learning of visual representations[C]//International Conference on Learning Representations: vol. 2. 2020: 4.

[208] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[J]. Advances in neural information processing systems, 2020, 33: 18661-18673.

[209] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.

[210] OH SONG H, XIANG Y, JEGELKA S, et al. Deep metric learning via lifted structured feature

embedding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4004-4012.

[211] SOHN K. Improved deep metric learning with multi-class n-pair loss objective[J]. Advances in neural information processing systems, 2016, 29.

[212] GE W. Deep metric learning with hierarchical triplet loss[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 269-285.

[213] YING Y, WEN L, LYU S. Stochastic online AUC maximization[J]. Advances in neural information processing systems, 2016, 29.

[214] LIU M, ZHANG X, CHEN Z, et al. Fast stochastic AUC maximization with $o(1/n)$-convergence rate[C]//International Conference on Machine Learning. 2018: 3189-3197.

[215] CLÉMENÇON S, LUGOSI G, VAYATIS N. Ranking and empirical minimization of U-statistics [J]. The Annals of Statistics, 2008: 844-874.

[216] AGARWAL S, NIYOGI P. Generalization bounds for ranking algorithms via algorithmic stability [J]. Journal of Machine Learning Research, 2009, 10(2).

[217] WANG Y, KHARDON R, PECHYONY D, et al. Generalization bounds for online learning algorithms with pairwise loss functions[C]//Conference on Learning Theory. 2012: 13-1.

[218] KAR P, SRIPERUMBUDUR B, JAIN P, et al. On the generalization ability of online learning algorithms for pairwise loss functions[C]//Proceedings of the 30th International Conference on Machine Learning. PMLR, 2013: 441-449.

[219] CAO Q, GUO Z C, YING Y. Generalization bounds for metric and similarity learning[J]. Machine Learning, 2016, 102(1): 115-132.

[220] LEI Y, LEDENT A, KLOFT M. Sharper generalization bounds for pairwise learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 21236-21246.

[221] LI S, LIU Y. Learning rates for nonconvex pairwise learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[222] WANG J, CHEN J, CHEN H, et al. Stability-based generalization analysis for mixtures of pointwise and pairwise learning[J]. arXiv preprint arXiv:2302.09967, 2023.

[223] HUANG S, ZHOU J, FENG H, et al. Generalization analysis of pairwise learning for ranking with deep neural networks[J]. Neural Computation, 2023: 1-24.

[224] CHEN J, CHEN H, JIANG X, et al. On the stability and generalization of triplet learning[J]. arXiv preprint arXiv:2302.09815, 2023.

[225] CHEN W, CHEN X, ZHANG J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 403-412.

[226] TANG H, LIU Y. Information-theoretic generalization bounds for transductive learning and its applications[J]. arXiv preprint arXiv:2311.04561, 2023.

[227] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.

[228] OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

[229] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. the Journal of machine Learning research, 2011, 12: 2825-2830.

[230] GEIRHOS R, RUBISCH P, MICHAELIS C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[C]//International Conference on Learning Representations. 2018.

[231] HENDRYCKS D, DIETTERICH T. Benchmarking neural network robustness to common corruptions and perturbations[C]//International Conference on Learning Representations. 2018.

[232] AZULAY A, WEISS Y. Why do deep convolutional networks generalize so poorly to small image transformations?[J]. Journal of Machine Learning Research, 2019, 20: 1-25.

[233] HENDRYCKS D, BASART S, MU N, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8340-8349.

[234] SUN B, SAENKO K. Deep coral: Correlation alignment for deep domain adaptation[C]//Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. 2016: 443-450.

[235] LI H, PAN S J, WANG S, et al. Domain generalization with adversarial feature learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5400-5409.

[236] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The journal of machine learning research, 2016, 17(1): 2096-2030.

[237] LI Y, TIAN X, GONG M, et al. Deep domain generalization via conditional invariant adversarial networks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 624-639.

[238] ARJOVSKY M, BOTTOU L, GULRAJANI I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019.

[239] CHEVALLEY M, BUNNE C, KRAUSE A, et al. Invariant causal mechanisms through distribution matching[J]. arXiv preprint arXiv:2206.11646, 2022.

[240] KOYAMA M, YAMAGUCHI S. When is invariance useful in an out-of-distribution generalization problem?[J]. arXiv preprint arXiv:2008.01883, 2020.

[241] SHI Y, SEELY J, TORR P, et al. Gradient matching for domain generalization[C]//International Conference on Learning Representations. 2021.

[242] RAME A, DANCETTE C, CORD M. Fishr: Invariant gradient variances for out-of-distribution generalization[C]//International Conference on Machine Learning. 2022: 18347-18377.

[243] SAGAWA S, KOH P W, HASHIMOTO T B, et al. Distributionally robust neural networks[C]//International Conference on Learning Representations. 2019.

[244] KRUEGER D, CABALLERO E, JACOBSEN J H, et al. Out-of-distribution generalization via risk extrapolation (rex)[C]//International Conference on Machine Learning. 2021: 5815-5826.

[245] EASTWOOD C, ROBEY A, SINGH S, et al. Probable domain generalization via quantile risk

minimization[J]. Advances in Neural Information Processing Systems, 2022, 35: 17340-17358.

[246] BLANCHARD G, DESHMUKH A A, DOGAN Ü, et al. Domain generalization by marginal transfer learning[J]. The Journal of Machine Learning Research, 2021, 22(1): 46-100.

[247] ZHANG M, MARKLUND H, DHAWAN N, et al. Adaptive risk minimization: Learning to adapt to domain shift[J]. Advances in Neural Information Processing Systems, 2021, 34: 23664-23678.

[248] NAGARAJAN V, ANDREASSEN A, NEYSHABUR B. Understanding the failure modes of out-of-distribution generalization[C]//International Conference on Learning Representations. 2020.

[249] PARASCANDOLO G, NEITZ A, ORVIETO A, et al. Learning explanations that are hard to vary [C]//International Conference on Learning Representations. 2020.

[250] GULRAJANI I, LOPEZ-PAZ D. In search of lost domain generalization[C]//International Conference on Learning Representations. 2020.

[251] MANSOUR Y, MOHRI M, ROSTAMIZADEH A. Domain adaptation: Learning bounds and algorithms[C]//22nd Conference on Learning Theory, COLT 2009. 2009.

[252] SHEN J, QU Y, ZHANG W, et al. Wasserstein distance guided representation learning for domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32. 2018.

[253] WANG Z, MAO Y. Information-theoretic analysis of unsupervised domain adaptation[C]//The Eleventh International Conference on Learning Representations. 2022.

[254] CHRISTIANSEN R, PFISTER N, JAKOBSEN M E, et al. A causal framework for distribution generalization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6614-6630.

[255] AHUJA K, CABALLERO E, ZHANG D, et al. Invariance principle meets information bottleneck for out-of-distribution generalization[J]. Advances in Neural Information Processing Systems, 2021, 34: 3438-3450.

[256] ZHOU X, LIN Y, ZHANG W, et al. Sparse invariant risk minimization[C]//International Conference on Machine Learning. 2022: 27222-27244.

[257] FEDERICI M, TOMIOKA R, FORRÉ P. An information-theoretic approach to distribution shifts [J]. Advances in Neural Information Processing Systems, 2021, 34: 17628-17641.

[258] LI B, WANG Y, ZHANG S, et al. Learning invariant representations and risks for semi-supervised domain adaptation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1104-1113.

[259] YING X. An overview of overfitting and its solutions[C]//Journal of physics: Conference series: vol. 1168. 2019: 022022.

[260] WANG X, SAXON M, LI J, et al. Causal balancing for domain generalization[C]//The Eleventh International Conference on Learning Representations. 2022.

[261] NGUYEN A T, TRAN T, GAL Y, et al. KL guided domain adaptation[C]//International Conference on Learning Representations. 2021.

[262] SHAHTALEBI S, GAGNON-AUDET J C, LALEH T, et al. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization[J]. arXiv preprint arXiv:2106.02266, 2021.

[263] KOLOURI S, NADJAHI K, SIMSEKLI U, et al. Generalized sliced Wasserstein distances[J]. Advances in neural information processing systems, 2019, 32.

[264] DESHPANDE I, HU Y T, SUN R, et al. Max-sliced Wasserstein distance and its use for gans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10648-10656.

[265] DAI B, SELJAK U. Sliced iterative normalizing flows[C]//International Conference on Machine Learning. 2021: 2352-2364.

[266] POOLADZANDI O, DAVINI D, MIRZASOLEIMAN B. Adaptive second order coresets for data-efficient machine learning[C]//International Conference on Machine Learning. 2022: 17848-17869.

[267] LIU Z, WANG Z, GUO H, et al. Over-training with mixup may hurt generalization[C]//The Eleventh International Conference on Learning Representations. 2024.

[268] DAS S. Inequalities for q-gamma function ratios[J]. Analysis and Mathematical Physics, 2019, 9(1): 313-321.

[269] XIANG S, CHEN X, WANG H. Error bounds for approximation in Chebyshev points[J]. Numerische Mathematik, 2010, 116(3): 463-491.

[270] LAM B. Some exact and asymptotic results for best uniform approximation.[D]. University of Tasmania, 1972.

[271] BERNSTEIN S. Collected Works: Constructive theory of functions (1905-1930)[M]. United States Atomic Energy Commission, 1952.

[272] BERNSTEIN S. Sur la meilleure approximation de $|x|^p$ par des polynômes de degrés très élevés[J]. Izv. Akad. Nauk SSSR Ser. Mat., 1938, 2(2): 169-190.

[273] VARGA R S, CARPENTER A J. Some numerical results on best uniform rational approximation of xα on [0,1][J]. Numerical Algorithms, 1992, 2(2): 171-185.

[274] KAWAGUCHI K, DENG Z, LUH K, et al. Robustness implies generalization via data-dependent generalization bounds[C]//International Conference on Machine Learning. 2022: 10866-10894.

附录 A 相关定理证明

A.1 信息论基础定义与引理

定义 附录 A.1 (次高斯性) 若随机变量 X 对任意 $\rho \in \mathbb{R}$ 均满足 $\mathbb{E}[\exp(\rho(X - \mathbb{E}[X]))] \leq \exp(\rho^2\sigma^2/2)$，则称其满足 σ-次高斯性。

定义 附录 A.2 (KL 散度) 设 P 与 Q 为定义在相同空间 $\mathcal{X}$ 上的概率分布，则 P 相对于 Q 的 KL 散度定义为 $D(P\,\|\,Q) \triangleq \int_{\mathcal{X}} p(x)\log(p(x)/q(x))\,\mathrm{d}x$。

定义 附录 A.3 (互信息) 设 (X, Y) 为一对定义在 $\mathcal{X} \times \mathcal{Y}$ 上的随机变量，其联合分布为 $P_{X,Y}$，边缘分布分别为 P_X 与 P_Y，则 X 与 Y 的互信息定义为 $I(X;Y) = D(P_{X,Y}\,\|\,P_XP_Y)$。

定义 附录 A.4 (Wasserstein 距离) 给定距离度量 $c(\cdot,\cdot)$，设 P 与 Q 为定义在 $\mathcal{X}$ 上的概率分布。记 $\Gamma(P,Q)$ 为 P 与 Q 所有耦合状态的集合（即 $\mathcal{X} \times \mathcal{X}$ 上所有边缘分布分别为 P 与 Q 的联合分布），则 P 与 Q 间的 p 阶 Wasserstein 距离定义为 $\mathbb{W}_p(P,Q) \triangleq \left(\inf_{\gamma\in\Gamma(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(x,x')^p\,\mathrm{d}\gamma(x,x')\right)^{1/p}$。

除非另外说明，以下默认使用 $\mathbb{W}(\cdot,\cdot)$ 表示一阶 Wasserstein 距离。

定义 附录 A.5 (全变差) 两个概率分布 P 与 Q 间的全变差定义为 $\mathrm{TV}(P,Q) \triangleq \sup_E |P(E) - Q(E)|$，其中最大值取于所有可测集 E。

定义 附录 A.6 (二值 KL 散度) 给定 $p, q \in [0,1]$，则 $d(q\,\|\,p)$ 表示参数为 q 与 p 的两个 Bernoulli 随机变量间的 KL 散度：$d(q\,\|\,p) = q\log(\frac{q}{p}) + (1-q)\log(\frac{1-q}{1-p})$。给定 $\gamma \in \mathbb{R}$，二值 KL 散度的一种松弛扩展定义为 $d_\gamma(q\,\|\,p) = \gamma q - \log(1 - p + pe^\gamma)$。容易验证，$\sup_\gamma d_\gamma(q\,\|\,p) = d(q\,\|\,p)$.

引理 附录 A.7 ([58]，引理 1) 设 (X, Y) 为一对具有联合分布 $P_{X,Y}$ 的随机变量，令 $\overline{Y}$ 为 Y 的一个独立拷贝。若函数 $f(x,y)$ 可测，$E_{X,Y}[f(X,Y)]$ 存在且 $f(X,\overline{Y})$ 满足 σ-次高斯性，则有

$$\left|\mathbb{E}_{X,Y}[f(X,Y)] - \mathbb{E}_{X,\overline{Y}}[f(X,\overline{Y})]\right| \leq \sqrt{2\sigma^2 I(X;Y)}. \tag{附录 A-1}$$

此外，若对于任意 x，$f(x,Y)$ 均满足 σ-次高斯性且以下期望存在，则有

$$\mathbb{E}_{X,Y}\left[\left(f(X,Y) - \mathbb{E}_{\overline{Y}}[f(X,\overline{Y})]\right)^2\right] \leq 4\sigma^2(I(X;Y) + \log 3), \tag{附录 A-2}$$

且对于任意 $\varepsilon > 0$，有

$$\Pr\left\{\left|f(X,Y) - \mathbb{E}_{\overline{Y}}[f(X,\overline{Y})]\right| \geq \varepsilon\right\} \leq \frac{4\sigma^2(I(X;Y) + \log 3)}{\varepsilon^2}. \tag{附录 A-3}$$

引理 附录 A.8 ([58]，引理 2) 设 X 满足 σ-次高斯性且 $\mathbb{E}[X] = 0$，则对于任意 $\lambda \in [0, 1/4\sigma^2)$，

$$\mathbb{E}_X\left[e^{\lambda X^2}\right] \leq 1 + 8\lambda\sigma^2. \tag{附录 A-4}$$

引理 附录 A.9 (Donsker-Varadhan 变分) 设 P 与 Q 为定义在相同可测空间 $\mathcal{X}$ 上的概率分布，其中 P 相对 Q 绝对连续。则对于任意有界可测函数 $f\colon \mathcal{X} \mapsto \mathbb{R}$，有

$$D(P\,\|\,Q) = \sup_f\left\{\mathbb{E}_{x\sim P}[f(x)] - \log \mathbb{E}_{x\sim Q}[e^{f(x)}]\right\}, \tag{附录 A-5}$$

其中 X 为任意使 e^X 满足 Q-可积性且 $\mathbb{E}_P[X]$ 存在的随机变量。

引理 附录 A.10 (Kantorovich-Rubinstein 对偶性) 设 P 与 Q 为定义在相同可测空间 $\mathcal{X}$ 上的概率分布，则

$$\mathbb{W}(P,Q) = \sup_{f\in \mathrm{Lip}_1}\left\{\int_{\mathcal{X}} f\mathrm{d}P - \int_{\mathcal{X}} f\mathrm{d}Q\right\}, \tag{附录 A-6}$$

其中 Lip_1 为度量 c 下满足 1-Lipschitz 连续性的函数集合，即对于任意 $f \in \mathrm{Lip}_1$ 与 $x, x' \in \mathcal{X}$，有 $|f(x) - f(x')| \leq c(x,x')$。

引理 附录 A.11 (Pinsker 不等式) 设 P 与 Q 为定义在相同空间上的概率分布，则

$$\mathrm{TV}(P,Q) \leq \sqrt{\frac{1}{2}D(Q\,\|\,P)}. \tag{附录 A-7}$$

引理 附录 A.12 ([51]，引理 2) 设 X 为几乎处处在 $[0,1]$ 范围内的随机变量且 $\mathbb{E}[X] = \mu$。则对于任意 $\gamma \in \mathbb{R}$，有

$$\mathbb{E}\left[e^{d_\gamma(X\,\|\,\mu)}\right] \leq 1. \tag{附录 A-8}$$

A.2 第 2 章定理补充证明

证明 (定理 2.3)**:** 设 $\{\psi_i\}_{i=1}^{N_\mathcal{H}}$ 为 $\mathcal{H}$ 的一组完全正交基。则根据 L'Hopital 法则，有

$$\begin{aligned}
\lim_{\alpha\to 1}\frac{1}{1-\alpha}\log \mathrm{tr}(G_X^\alpha) &= \lim_{\alpha\to 1} -\frac{\frac{\partial}{\partial\alpha}\mathrm{tr}(G_X^\alpha)}{\mathrm{tr}(G_X^\alpha)} = -\mathrm{tr}(G_X\log G_X)\\
&= -\sum_{i=1}^{N_\mathcal{H}}\langle G_X\psi_i, \log G_X\psi_i\rangle\\
&= -\sum_{i=1}^{N_\mathcal{H}}\int_{\mathcal{X}} p_X(x)\langle\psi_i,\varphi(x)\rangle\langle\varphi(x),\log G_X\psi_i\rangle\,\mathrm{d}x\\
&= -\int_{\mathcal{X}} p_X(x)\langle\log G_X\varphi(x), \sum_{i=1}^{N_\mathcal{H}}\langle\psi_i,\varphi(x)\rangle\psi_i\rangle\,\mathrm{d}x\\
&= -\int_{\mathcal{X}} p_X(x)\langle\log G_X\varphi(x),\varphi(x)\rangle\,\mathrm{d}x\\
&= -\int_{\mathcal{X}} p_X(x)\langle\int_{\mathcal{X}}\log p_X(x')\varphi(x')\langle\varphi(x'),\varphi(x)\rangle\,\mathrm{d}x',\varphi(x)\rangle\,\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= -\iint_{\mathcal{X}^2} p_X(x)\log p_X(x')\langle\varphi(x'),\varphi(x)\rangle\langle\varphi(x'),\varphi(x)\rangle\,\mathrm{d}x\,\mathrm{d}x' \\
&= -\iint_{\mathcal{X}^2} p_X(x)\log p_X(x')\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x'. \qquad \blacksquare
\end{aligned}
$$

证明 (定理 2.5)**:** 设 λ_i 与 μ_i 分别为 G_X 与 $\widehat{G}_X$ 的特征值。应用 [185] 中定理 6.2 的证明方法，并取 $\phi(x)=|x|$，则以概率 $1-\delta$，有

$$
\sum_{i=1}^{m}|\lambda_i-\mu_i| \le C\sqrt{\frac{2\log\frac{2}{\delta}}{m}}, \tag{附录 A-9}
$$

其中 $C=\max_{x\in\mathcal{X}}\kappa(x,x)=1$。对于任意 $s>0$，有

$$
\begin{aligned}
&\left|\mathrm{tr}(G_X\log G_X)-\mathrm{tr}(\widehat{G}_X\log\widehat{G}_X)\right| \\
&= \left|\sum_{i=1}^{m}\lambda_i\log\lambda_i-\sum_{i=1}^{m}\mu_i\log\mu_i\right| \le \sum_{i=1}^{m}|\lambda_i\log\lambda_i-\mu_i\log\mu_i| \\
&\le \sum_{i=1}^{m}\max\Big(-|\lambda_i-\mu_i|\log|\lambda_i-\mu_i|,-(1-|\lambda_i-\mu_i|)\log(1-|\lambda_i-\mu_i|)\Big) \qquad \text{(附录 A-10)} \\
&\le \sum_{i=1}^{m}-\sqrt{\frac{2\log\frac{2}{\delta}}{m^3}}\log\sqrt{\frac{2\log\frac{2}{\delta}}{m^3}} = \sqrt{\frac{2\log\frac{2}{\delta}}{m}}\log\sqrt{\frac{m^3}{2\log\frac{2}{\delta}}} \\
&\le s\sqrt{\frac{2\log\frac{2}{\delta}}{m}}\left(\frac{m^3}{2\log\frac{2}{\delta}}\right)^{\frac{1}{2s}} \le sm^{\frac{3}{2s}-\frac{1}{2}}\sqrt{2\log\frac{2}{\delta}}, \qquad \text{(附录 A-11)}
\end{aligned}
$$

其中式 (附录 A-10) 在 $|\lambda_1-\mu_1|=\cdots=|\lambda_n-\mu_n|=\frac{1}{m}\sqrt{\frac{2\log\frac{2}{\delta}}{m}}$ 时达到最大值，式 (附录 A-11) 可由以下事实得出：对任意 $t>0$，有 $\log x\le x^t/t$。当取 $s=9$ 时，有

$$
\begin{aligned}
|S_1(X)-\widehat{S}_1(X)| &= C_\kappa\left|\mathrm{tr}(G_X\log G_X)-\mathrm{tr}(K\log K)\right| \\
&= C_\kappa\left|\mathrm{tr}(G_X\log G_X)-\mathrm{tr}(\widehat{G}_X\log\widehat{G}_X)\right| \le \frac{9C_\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt[3]{m}}. \qquad \blacksquare
\end{aligned}
$$

证明 (定理 2.8)**:** 对任意定义在 $\mathcal{X}$ 上的概率密度函数 $p(\cdot)$：

$$
\begin{aligned}
\lim_{c\to0}E_X^\kappa(p) &= \lim_{c\to0}C_\kappa\iint_{\mathcal{X}^2}p(x)\Big[\log p_X(x)-\log p_X(x')\Big]\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= \lim_{c\to0}C_\kappa\iint_{\mathcal{X}^2}p(x)\Big[\log p_X(x)-\log p_X(x+x')\Big]\kappa^2(0,x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= C_\kappa\int_{\mathcal{X}}\left\{\lim_{c\to0}\int_{\|x-x'\|<c}p(x)\Big[\log p_X(x)-\log p_X(x+x')\Big]\,\mathrm{d}x\right\}\mathrm{d}x' = C_\kappa\int_{\mathcal{X}}0\,\mathrm{d}x' = 0. \blacksquare
\end{aligned}
$$

证明 (定理 2.9，性质 1)**:**

$$
\begin{aligned}
S_1(X) - H(X) &= \int_{\mathcal{X}} p_X(x)\log p_X(x)\,\mathrm{d}x - C_\kappa \iint_{\mathcal{X}^2} p_X(x)\log p_X(x')\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= \int_{\mathcal{X}} p_X(x)\log p_X(x)\,\mathrm{d}x - \int_{\mathcal{X}} p_X(x)\left(C_\kappa \int_{\mathcal{X}} \log p_X(x')\kappa^2(x,x')\,\mathrm{d}x'\right)\mathrm{d}x \\
&\geq \int_{\mathcal{X}} p_X(x)\log p_X(x)\,\mathrm{d}x - \int_{\mathcal{X}} p_X(x)\left(\log C_\kappa \int_{\mathcal{X}} p_X(x')\kappa^2(x,x')\,\mathrm{d}x'\right)\mathrm{d}x \qquad \text{(附录 A-12)} \\
&= \int_{\mathcal{X}} p_X(x)\log \frac{p_X(x)}{C_\kappa \int_{\mathcal{X}} p_X(x')\kappa^2(x,x')\,\mathrm{d}x'}\,\mathrm{d}x = D(P_X \,\|\, Q_X) \geq 0,
\end{aligned}
$$

其中式 (附录 A-12) 可由 Jensen 不等式得出，且 Q_X 为具有概率密度

$$
q_X(x) = C_\kappa \int_{\mathcal{X}} p_X(x')\kappa^2(x,x')\,\mathrm{d}x', \qquad \text{(附录 A-13)}
$$

的概率分布。据此可得

$$
\begin{aligned}
S_1(X) &= -C_\kappa \iint_{\mathcal{X}^2} p_X(x)\log p_X(x')\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' = -\int_{\mathcal{X}} p_X(x)\left(C_\kappa \int_{\mathcal{X}} \log p_X(x')\kappa^2(x,x')\,\mathrm{d}x'\right)\mathrm{d}x \\
&= -\int_{\mathcal{X}} p_X(x)\log p_X(x)\,\mathrm{d}x - \int_{\mathcal{X}} p_X(x)\left[C_\kappa \int_{\mathcal{X}} \Big(\log p_X(x') - \log p_X(x)\Big)\kappa^2(x,x')\,\mathrm{d}x'\right]\mathrm{d}x \\
&\leq H(X) + E_X^{\kappa\prime}. \qquad \blacksquare
\end{aligned}
$$

证明 (定理 2.9，性质 2)**:** 设 p 与 q 分别为 X 与 X' 的概率密度函数。考虑以下 $q(x)$ 为输入的函数：

$$
J(q) = C_\kappa \iint_{\mathcal{X}^2} p(x)\log \frac{p(x')}{q(x')}\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' - \eta_0\left(\int_{\mathcal{X}} q(x)\,\mathrm{d}x - 1\right),
$$

其中 η_0 为 Lagrange 乘子，其目的为确保 $q(x)$ 可构成概率密度。散度 $D_1(P \,\|\, Q)$ 在以下导数为 0 时达到最大值：

$$
\frac{\partial J}{\partial q} = -C_\kappa \int_{\mathcal{X}} \frac{p(x)}{q(x')}\kappa^2(x,x')\,\mathrm{d}x - \eta_0 = -\frac{C_\kappa \int_{\mathcal{X}} p(x)\kappa^2(x,x')\,\mathrm{d}x}{q(x')} - \eta_0 = 0,
$$

这表明极小值 $\hat{q}(x')$ 满足 $\hat{q}(x') \propto C_\kappa \int_{\mathcal{X}} p(x)\kappa^2(x,x')\,\mathrm{d}x$。结合 $\int_{\mathcal{X}} \hat{q}(x')\,\mathrm{d}x' = 1$，有

$$
\hat{q}(x') = C_\kappa \int_{\mathcal{X}} p(x)\kappa^2(x,x')\,\mathrm{d}x.
$$

$$
\begin{aligned}
D_1(P \,\|\, Q) &\geq C_\kappa \iint_{\mathcal{X}^2} p(x)\log \frac{p(x')}{\hat{q}(x')}\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= \int_{\mathcal{X}} \left(C_\kappa \int_{\mathcal{X}} p(x)\kappa^2(x,x')\,\mathrm{d}x\right)\log \frac{p(x')}{\hat{q}(x')}\,\mathrm{d}x' \\
&\geq -\int_{\mathcal{X}} \hat{q}(x')\left(C_\kappa \int_{\mathcal{X}} (\log p(x') - \log p(x))\kappa^2(x,x')\,\mathrm{d}x\right)\mathrm{d}x' \geq -E_X^{\kappa}.
\end{aligned}
$$

上式中最后一步可由 Jensen 不等式得出。 $\blacksquare$

证明 (定理 2.9，性质 3)**:** 首个等式可直接通过相关定义（定义 2.6 与定义 2.7）得出。$I_1(X;Y)$ 的正值性可在 [185]，命题 4.1 中置 $n \to \infty$ 得到。 ■

证明 (定理 2.9，性质 4)**:** 注意到

$$\begin{aligned} S_1(X) &= -C_{\kappa_X} \iint_{\mathcal{X}^2} p_X(x) \log p_X(x') \kappa_X^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \\ &= -C_{\kappa_X} \iiint_{\mathcal{Y}\times\mathcal{X}^2} p_{X,Y}(x,y) \log p_X(x') \kappa_X^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \\ &= -C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2\times\mathcal{X}^2} p_{X,Y}(x,y) \log p_X(x') \kappa_X^2(x,x') \kappa_Y^2(y,y') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \,\mathrm{d}y'. \end{aligned}$$

类似地，有 $S_1(Y) = -C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2\times\mathcal{X}^2} p_{X,Y}(x,y) \log p_Y(y') \kappa_X^2(x,x') \kappa_Y^2(y,y') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \,\mathrm{d}y'$。结合以上等式即可完成证明。 ■

证明 (定理 2.9，性质 5)**:** 根据核化 Rényi 互信息的定义，可进一步定义核化 Rényi 条件互信息：

$$\begin{aligned} I_1(X;Y|Z) = C_{\kappa_X} C_{\kappa_Y} C_{\kappa_Z} \iint_{\mathcal{Z}^2} \iint_{\mathcal{Y}^2} \iint_{\mathcal{X}^2} p_{X,Y,Z}(x,y,z) \log \frac{p_{X,Y|Z}(x',y'|z')}{p_{X|Z}(x'|z') p_{Y|Z}(y'|z')} \\ \cdot \kappa_X^2(x,x') \kappa_Y^2(y,y') \kappa_Z^2(z,z') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \,\mathrm{d}y' \,\mathrm{d}z \,\mathrm{d}z'. \end{aligned} \tag{附录 A-14}$$

本性质可直接通过上述定义得出。 ■

证明 (定理 2.9，性质 6)**:** 根据性质 5，有

$$I_1(X;Y,Z) = I_1(X;Y|Z) + I_1(X;Z) = I_1(X;Z|Y) + I_1(X;Y).$$

根据 Markov 条件，X 与 Z 在给定 Y 时条件独立，即 $p_{X,Z|Y}(x,z|y) = p_{X|Y}(x|y) p_{Z|Y}(z|y)$。根据条件互信息的定义，有 $I_1(X;Z|Y) = 0$。则本性质的首个不等式可通过 $I_1(X;Y|Z)$ 的非负性得出。类似地，可证明本性质的第二个不等式。 ■

引理 附录 A.13 设 P 与 Q 为定义在相同空间上的概率分布，其中 P 相对于 Q 满足绝对连续性。则

$$D_1(P \parallel Q) + E_P^\kappa \geq \mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X].$$

其中 X 为任意使 e^X 满足 Q-可积性且 $\mathbb{E}_P[X]$ 存在的随机变量。

证明: 定义 Q^X 为满足以下性质的概率分布：

$$Q^X(\Omega) = \int_\Omega \frac{e^X}{\mathbb{E}_Q[e^X]} \,\mathrm{d}Q,$$

则 Q 相对于 Q^X 绝对连续。观察到

$$D_1(P \parallel Q) + E_P^\kappa = D_1(P \parallel Q^X) + E_P^\kappa + C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log \frac{e^{x'}}{\mathbb{E}_Q[e^X]} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x'$$

$$
\begin{aligned}
&\geq C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log e^{x'} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' - C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log \mathbb{E}_Q[e^X] \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \\
&= \int_{\mathcal{X}} p_X(x) \left(C_\kappa \int_{\mathcal{X}} x' \kappa^2(x,x') \,\mathrm{d}x' \right) \mathrm{d}x - \log \mathbb{E}_Q[e^X] \int_{\mathcal{X}} p_X(x) \left(C_\kappa \int_{\mathcal{X}} \kappa^2(x,x') \,\mathrm{d}x' \right) \mathrm{d}x \\
&= \int_{\mathcal{X}} p_X(x) x \,\mathrm{d}x - \log \mathbb{E}_Q[e^X] \int_{\mathcal{X}} p_X(x) \,\mathrm{d}x = \mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X]. \qquad \blacksquare
\end{aligned}
$$

引理 附录 A.14 ([58]，引理 3) 设 X 与 Y 为独立随机变量，若可测函数 f 使对于任意 $x \in \mathcal{X}$，$f(x,Y)$ 均满足 σ-次高斯性且 $\mathbb{E}_Y[f(x,Y)] = 0$，则 $f(X,Y)$ 同样满足 σ-次高斯性。

证明 (定理 2.10)**:** 设 $f(w,z) = L(w) - L_z(w)$，W' 与 $\mathbf{Z}'$ 分别为 W 与 $\mathbf{Z}$ 的独立拷贝，则对于任意 $\lambda \in [0,\infty)$，有

$$
\begin{aligned}
I_1(W;\mathbf{Z}) + E^\kappa_{W,\mathbf{Z}} &= D_1(P_{W,\mathbf{Z}} \parallel P_W \otimes P_{\mathbf{Z}}) + E^\kappa_{W,\mathbf{Z}} \\
&\geq \mathbb{E}_{W,\mathbf{Z}}[\lambda f(W,\mathbf{Z})] - \log \mathbb{E}_{W',\mathbf{Z}'} \left[e^{\lambda f(W',\mathbf{Z}')} \right] \\
&= \mathbb{E}_{W,\mathbf{Z}}[\lambda f(W,\mathbf{Z})] - \log \mathbb{E}_{W,\mathbf{Z}'} \left[e^{\lambda f(W,\mathbf{Z}')} \right] \qquad \text{(附录 A-15)}
\end{aligned}
$$

上式可通过引理 附录 A.13 以及 $W, W', \mathbf{Z}'$ 的独立性得出。注意到 $f(w,\mathbf{Z})$ 对于任意 $w \in \mathcal{W}$ 均满足 $\sigma/\sqrt{n}$-次高斯性：$L_{\mathbf{Z}}(w)$ 是 n 个独立同分布 σ-次高斯随机变量的平均值。此外，对于任意 w 有 $\mathbb{E}_{\mathbf{Z}}[f(w,\mathbf{Z})] = 0$。应用引理 附录 A.14，可知 $f(W,\mathbf{Z})$ 满足 $\sigma/\sqrt{n}$-次高斯性。因此，有

$$
\begin{aligned}
\log \mathbb{E}_{W,\mathbf{Z}'} \left[e^{\lambda f(W,\mathbf{Z}') - \lambda \mathbb{E}_{W,\mathbf{Z}'}[f(W,\mathbf{Z}')]} \right] &\leq \frac{\lambda^2 \sigma^2}{2n}, \\
\log \mathbb{E}_{W,\mathbf{Z}'} \left[e^{\lambda f(W,\mathbf{Z}')} \right] &\leq \frac{\lambda^2 \sigma^2}{2n}.
\end{aligned}
$$

代入式 (附录 A-15) 中，可得

$$
I_1(W;\mathbf{Z}) + E^\kappa_{W,\mathbf{Z}} \geq \lambda \mathbb{E}_{W,\mathbf{Z}}[f(W,\mathbf{Z})] - \frac{\lambda^2 \sigma^2}{2} \geq \frac{n}{2\sigma^2} \mathbb{E}^2_{W,\mathbf{Z}}[f(W,\mathbf{Z})].
$$

第一部分得证。对于第二部分，设 $\tilde{f}(w,z) = (L(w) - L_z(w))^2$，且 $\lambda \in \left[0, \frac{1}{4\sigma^2}\right)$，则有

$$
\begin{aligned}
I_1(W;\mathbf{Z}) + E^\kappa_{W,\mathbf{Z}} &\geq \mathbb{E}_{W,\mathbf{Z}} \left[\lambda \tilde{f}(W,\mathbf{Z}) \right] - \log \mathbb{E}_{W',\mathbf{Z}'} \left[e^{\lambda \tilde{f}(W',\mathbf{Z}')} \right] \\
&\geq \mathbb{E}_{W,\mathbf{Z}} \left[\lambda (L(W) - L_{\mathbf{Z}}(W))^2 \right] - \log(1 + 8\lambda\sigma^2) \\
&\geq \frac{1}{4\sigma^2} \mathbb{E}_{W,\mathbf{Z}} \left[(L(W) - L_{\mathbf{Z}}(W))^2 \right] - \log 3,
\end{aligned}
$$

上式可通过在引理 附录 A.13 以及引理 附录 A.8 中取 $\lambda = \frac{1}{4\sigma^2}$ 得到。第二部分得证。 $\blacksquare$

引理 附录 A.15 设连续型随机变量 $X \sim N(0,\Sigma)$ 与 X' 满足 $\mathbb{E}[X'] = 0$ 与 $\mathrm{Cov}[X'] = \Sigma$，则

$S_1(X') \le S_1(X) + E^{\kappa}_{X'}$。另外，若 κ 是宽度为 σ_κ 的高斯核函数，则 $S_1(X) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log|\Sigma| + \frac{\sigma_\kappa^2}{4}\mathrm{tr}[\Sigma^{-1}]$.

证明： 设 $p(\cdot)$ 与 $q(\cdot)$ 分别为 X 与 X' 的概率密度函数。注意到 $p_\kappa = C_\kappa \kappa^2(0,\cdot)$ 积分为 1，因此可作为概率密度函数，其对应协方差矩阵为 $\frac{1}{2}\sigma_\kappa^2 I$。则有

$$
\begin{aligned}
S_1(X) - S_1(X') &= C_\kappa \iint_{\mathcal{X}^2} \left[q(x)\left(\log p(x') + \log \frac{q(x')}{p(x')} \right) - p(x)\log p(x') \right] \kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= C_\kappa \iint_{\mathcal{X}^2} [q(x)-p(x)] \log p(x') \kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' + C_\kappa \iint_{\mathcal{X}^2} q(x)\log\frac{q(x')}{p(x')}\kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' \\
&= C_\kappa \iint_{\mathcal{X}^2} [q(x)-p(x)] \left(-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}x'^{\top}\Sigma^{-1}x' \right) \kappa^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' + D_1(Q \,\|\, P) \\
&\ge -\left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \right) \int_{\mathcal{X}} [q(x)-p(x)] \left(C_\kappa \int_{\mathcal{X}} \kappa^2(x,x')\,\mathrm{d}x' \right) \mathrm{d}x \\
&\quad - \frac{1}{2}\mathrm{tr}\left\{ \left[\int_{\mathcal{X}} [q(x)-p(x)] \left(C_\kappa \int_{\mathcal{X}} x' x'^{\top} \kappa^2(x,x')\,\mathrm{d}x' \right) \mathrm{d}x \right] \Sigma^{-1} \right\} - E^{\kappa}_{X'} \\
&= -\left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \right) \int_{\mathcal{X}} [q(x)-p(x)]\,\mathrm{d}x \\
&\quad - \frac{1}{2}\mathrm{tr}\left\{ \left[\int_{\mathcal{X}} [q(x)-p(x)] \left(xx^{\top} + \frac{1}{2}\sigma_\kappa^2 I \right) \mathrm{d}x \right] \Sigma^{-1} \right\} - E^{\kappa}_{X'} \\
&= -\frac{1}{2}\mathrm{tr}\left[\left(\int_{\mathcal{X}} [q(x)-p(x)] xx^{\top}\,\mathrm{d}x \right) \Sigma^{-1} \right] - E^{\kappa}_{X'} = -\frac{1}{2}\mathrm{tr}\left[(\Sigma - \Sigma)\Sigma^{-1} \right] - E^{\kappa}_{X'} = -E^{\kappa}_{X'}.
\end{aligned}
$$

考虑 κ 为高斯核函数的情形，即 $\kappa^2(x,x') = \exp\left(-\|x-x'\|_2^2/\sigma_\kappa^2\right)$ 且 $C_\kappa = (\pi\sigma_\kappa^2)^{-d/2}$，则有

$$
\begin{aligned}
S_1(X) &= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \int_{\mathcal{X}} \frac{1}{2}\exp\left(-\frac{1}{2}x^{\top}\Sigma^{-1}x \right) \mathrm{tr}\left[\left(xx^{\top} + \frac{1}{2}\sigma_\kappa^2 I \right)\Sigma^{-1} \right] \mathrm{d}x \\
&= \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{1}{2}\mathrm{tr}\left[\left(\Sigma + \frac{1}{2}\sigma_\kappa^2 I \right)\Sigma^{-1} \right] = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log|\Sigma| + \frac{\sigma_\kappa^2}{4}\mathrm{tr}[\Sigma^{-1}]. \qquad \blacksquare
\end{aligned}
$$

引理 附录 A.16 设 X、Y、Δ 与 $\xi \sim N(0,\sigma^2 I)$ 为独立随机变量，给定函数 $f\colon \mathcal{W}\times\mathcal{Z}^b \to \mathcal{W}$ 并设 $\Omega(\cdot) = \mathbb{E}_X[f(\cdot, X)]$，则有

$$
I_1(f(Y+\Delta, X) + \xi; X|Y) \le \frac{1}{2}\log\left| \frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)] + I \right| + E^{\kappa}_{f(Y+\Delta,X)-\Omega(Y+\Delta)+\xi|Y,\Delta}.
$$

证明： 设 $g(Y,\Delta,X) = f(Y+\Delta,X) - \Omega(Y+\Delta)$，则有

$$
\begin{aligned}
&I_1(f(Y+\Delta,X)+\xi; X|Y=y,\Delta=\delta) \qquad \text{(附录 A-16)} \\
&= I_1(g(Y,\Delta,X)+\xi; X|Y=y,\Delta=\delta)
\end{aligned}
$$

$$
\begin{aligned}
&= S_1(g(Y,\Delta,X)+\xi|Y=y,\Delta=\delta) - S_1(g(Y,\Delta,X)+\xi|X,Y=y,\Delta=\delta)\\
&= S_1(g(Y,\Delta,X)+\xi|Y=y,\Delta=\delta) - S_1(\xi)\\
&= S_1(g(Y,\Delta,X)+\xi|Y=y,\Delta=\delta) - \frac{d}{2}\log(2\pi e\sigma^2) - \frac{d\sigma_\kappa^2}{4\sigma^2}\\
&\le \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\left|\mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta]+\sigma^2 I\right| - \frac{d}{2}\log(2\pi e\sigma^2) + E^{\kappa}_{g(Y,\Delta,X)+\xi|Y,\Delta}\\
&\quad + \frac{\sigma_\kappa^2}{4}\mathrm{tr}\left[\left(\mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta]+\sigma^2 I\right)^{-1}\right] - \frac{d\sigma_\kappa^2}{4\sigma^2} \qquad \text{(附录 A-17)}\\
&\le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta]+I\right| + E^{\kappa}_{g(Y,\Delta,X)+\xi|Y,\Delta} \qquad \text{(附录 A-18)}
\end{aligned}
$$

其中式 (附录 A-17) 可通过引理 附录 A.15 及下式得到：

$$
\begin{aligned}
\mathrm{Cov}[g(Y,\Delta,X)+\xi|Y=y,\Delta=\delta] &= \mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta] + \mathrm{Cov}[\xi]\\
&= \mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta] + \sigma^2 I,
\end{aligned}
$$

式 (附录 A-18) 可通过下式得到：

$$
\mathrm{tr}\left[\left(\mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta]+\sigma^2 I\right)^{-1}\right] \le \mathrm{tr}\left[\left(\sigma^2 I\right)^{-1}\right].
$$

继而可得如下的互信息上界：

$$
\begin{aligned}
&I_1(f(Y+\Delta,X)+\xi;X|Y)\\
&\le I_1(f(Y+\Delta,X)+\xi,\Delta;X|Y)\\
&= I_1(f(Y+\Delta,X)+\xi,\Delta;X|Y) - I_1(\Delta;X|Y) \qquad \text{(附录 A-19)}\\
&= I_1(f(Y+\Delta,X)+\xi;X|Y,\Delta)\\
&= \mathbb{E}_{Y,\Delta}[I_1(f(Y+\Delta,X)+\xi;X|Y=y,\Delta=\delta)]\\
&\le \mathbb{E}_{Y,\Delta}\left[\frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta]+I\right|\right] + E^{\kappa}_{g(Y,\Delta,X)+\xi|Y,\Delta} \qquad \text{(附录 A-20)}\\
&\le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathbb{E}_{Y,\Delta}[\mathrm{Cov}[g(Y,\Delta,X)|Y=y,\Delta=\delta]]+I\right| + E^{\kappa}_{g(Y,\Delta,X)+\xi|Y,\Delta} \qquad \text{(附录 A-21)}\\
&= \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)]+I\right| + E^{\kappa}_{g(Y,\Delta,X)+\xi|Y,\Delta}, \qquad \text{(附录 A-22)}
\end{aligned}
$$

其中式 (附录 A-19) 可由 Δ 与 X 的独立性得到，式 (附录 A-20) 可由式 (附录 A-18) 得到，式 (附录 A-21) 可由 Jensen 不等式及对数行列式函数的凹性得到，式 (附录 A-22) 可通过全方差公式得到。 ■

证明 (定理 2.13)**：** 注意到 $\mathbf{Z}\to(B_1,\cdots,B_T)\to(W_1,\cdots,W_T)$ 构成 Markov 链，则有

$$\begin{aligned}I_1(W_T;\mathbf{Z}) &\le I_1(W_T;B_1,\cdots,B_T) \le I_1(W_0,W_1,\cdots,W_T;B_1,\cdots,B_T)\\ &= I_1(W_0;B_1,\cdots,B_T)+I_1(W_1;B_1,\cdots,B_T|W_0)+I_1(W_2;B_1,\cdots,B_T|W_0,W_1)\\ &\quad+\cdots+I_1(W_T;B_1,\cdots,B_T|W_0,\cdots,W_T).\end{aligned}$$

对于任意 $t\in[1,T]$，有

$$\begin{aligned}I_1(W_t;B_1,\cdots,B_t|W_0,\cdots,W_{t-1}) &= S_1(W_t|W_0,\cdots,W_{t-1})-S_1(W_t|B_1,\cdots,B_t,W_0,\cdots,W_{t-1})\\ &= S_1(W_t|W_{t-1})-S_1(W_t|B_t,W_{t-1})\\ &= I_1(W_t;B_t|W_{t-1}).\end{aligned}$$

最后，在引理 附录 A.16 中取 $X=B_t$，$Y=W_{t-1}$，$\Delta=0$，$\xi=\xi_t$ 以及 $f(W_{t-1},B_t)=-\eta_t g(W_{t-1},B_t)$，可得

$$\begin{aligned}I_1(W_t;B_t|W_{t-1}) &= I_1(W_t-W_{t-1};B_t|W_{t-1}) = I_1(-\eta_t g(W_{t-1},B_t)+\xi_t;B_t|W_{t-1})\\ &\le \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathrm{Cov}[g(W_{t-1},B_t)]+I\right|+E^{\kappa}_{W_t|W_{t-1}},\end{aligned}$$ ■

引理 附录 A.17 设 V 为 $n\times n$ 对称正定矩阵，并可划分为

$$V=\begin{bmatrix}A & C^\top\\ C & B\end{bmatrix},$$

其中 A、B 是大小分别为 $n_1\times n_1$ 与 $n_2\times n_2$ 的对称矩阵，则有 $|V|\le|A||B|$。

证明: 注意到

$$V=D\begin{bmatrix}A & 0\\ 0 & B-CA^{-1}C^\top\end{bmatrix}D^\top,\quad \text{其中 } D=\begin{bmatrix}I_{n_1} & 0\\ CA^{-1} & I_{n_2}\end{bmatrix},$$

则有 $|V|=|D||A||B-CA^{-1}C^\top||D^\top|=|A||B-CA^{-1}C^\top|$。设 $D^\dagger$ 为 D 的伪逆，则对于任意长度为 n 的列向量 x：

$$x^\top\begin{bmatrix}A & 0\\ 0 & B-CA^{-1}C^\top\end{bmatrix}x=(x^\top D^\dagger)V(x^\top D^\dagger)^\top\ge 0,$$

因此 $B-CA^{-1}C^\top$ 为半正定矩阵。类似地，可证明 $CA^{-1}C^\top$ 为半正定矩阵。设 λ_i、μ_i 与 v_i，$i\in\{1,\cdots,n_2\}$ 分别为矩阵 $B-CA^{-1}C^\top$、B 与 $CA^{-1}C^\top$ 从大到小排列的特征值，则通过 Weyl 不等式，对于任意 $i\in\{1,\cdots,n_2\}$，有 $\lambda_i\le\mu_i-v_{n_2}\le\mu_i$，表明 $|B-CA^{-1}C^\top|\le|B|$。综合以上结果，定理得证：$|V|=|A||B-CA^{-1}C^\top|\le|A||B|$. ■

证明 (定理 2.14)**:** 注意到 $V_t = \mathrm{tr}[\mathcal{V}_t]$。由于协方差矩阵总是满足对称正定性，故可将其特征值表示为 λ_1、$\cdots$、$\lambda_d \geq 0$。注意到

$$\begin{aligned}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t + I\right| &= \log\left[\prod_{i=1}^{d}\left(\frac{\eta_t^2\lambda_i}{\sigma_t^2}+1\right)\right] \leq \log\left[\frac{1}{d}\sum_{i=1}^{d}\left(\frac{\eta_t^2\lambda_i}{\sigma_t^2}+1\right)\right]^d \\ &= d\log\left[\frac{\eta_t^2}{d\sigma_t^2}\sum_{i=1}^{d}\lambda_i + 1\right] = d\log\left[\frac{\eta_t^2 V_t}{d\sigma_t^2}+1\right],\end{aligned}$$

上式中的不等式可通过几何平均数总是小于算术平均数的性质得到。设 $V_t^i = \mathrm{tr}[\mathcal{V}_t^i]$，可通过类似方法证明，对于任意 $i \in \{1, \cdots, r\}$ 均有

$$\theta_c(\mathcal{V}_t^i) \leq \theta_v(V_t^i),$$

通过 Jensen 不等式，有

$$\begin{aligned}\sum_{i=1}^{r}\theta_c(\mathcal{V}_t^i) \leq \sum_{i=1}^{r}\theta_v(V_t^i) &= \sum_{i=1}^{r} d\log\left[\frac{\eta_t^2 V_t^i}{d\sigma_t^2}+1\right] \leq d\log\left[\frac{\eta_t^2}{d\sigma_t^2}\sum_{i=1}^{r}V_t^i + 1\right] \\ &= d\log\left[\frac{\eta_t^2}{d\sigma_t^2}V_t + 1\right] = \theta_v(V_t).\end{aligned}$$

通过递归应用引理 附录 A.17，可证明

$$\theta_c(\mathcal{V}_t) = \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t + I\right| \leq \frac{1}{2}\log\prod_{i=1}^{r}\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t^i + I\right| = \frac{1}{2}\sum_{i=1}^{r}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathcal{V}_t^i + I\right| = \sum_{i=1}^{r}\theta_c(\mathcal{V}_t^i).$$

对于定理中的最后一个不等式，注意到

$$\begin{aligned}V_t &= \mathbb{E}_{B_t}[\|g(W_{t-1},B_t) - \mathbb{E}_{B_t}[g(W_{t-1},B_t)]\|_2^2] = \mathbb{E}_{B_t}[\|g(W_{t-1},B_t)\|_2^2] - \|\mathbb{E}_{B_t}[g(W_{t-1},B_t)]\|_2^2 \\ &\leq \mathbb{E}_{B_t}[\|g(W_{t-1},B_t)\|_2^2] \leq \max_{w\in\mathcal{W},z\in\mathcal{Z}}\|g(w,z)\|_2^2 = L,\end{aligned}$$

通过 log 函数的单调性，定理得证。 ■

引理 附录 A.18 互信息 $I_1(\widetilde{W}_T;\mathbf{Z}) \leq \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t) + \widetilde{\xi}_t; B_t|\widetilde{W}_{t-1})$.

证明:

$$\begin{aligned}I_1(\widetilde{W}_T;\mathbf{Z}) &= I_1(\widetilde{W}_{T-1} - \eta_T g(W_{T-1},B_T) + \widetilde{\xi}_T;\mathbf{Z}) \\ &\leq I_1(\widetilde{W}_{T-1}, -\eta_T g(W_{T-1},B_T) + \widetilde{\xi}_T;\mathbf{Z}) && \text{(附录 A-23)} \\ &= I_1(\widetilde{W}_{T-1};\mathbf{Z}) + I_1(-\eta_T g(W_{T-1},B_T) + \widetilde{\xi}_T;\mathbf{Z}|\widetilde{W}_{T-1}) && \text{(附录 A-24)} \\ &\leq I_1(\widetilde{W}_{T-2};\mathbf{Z}) + I_1(-\eta_{T-1} g(W_{T-2},B_{T-1}) + \widetilde{\xi}_{T-1};\mathbf{Z}|\widetilde{W}_{T-2}) \\ &\quad + I_1(-\eta_T g(W_{T-1},B_T) + \widetilde{\xi}_T;\mathbf{Z}|\widetilde{W}_{T-1})\end{aligned}$$

$$
\begin{aligned}
&\leq \cdots \\
&\leq I_1(\widetilde{W}_0;\mathbf{Z}) + \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t) + \widetilde{\xi}_t;\mathbf{Z}|\widetilde{W}_{t-1}) && \text{(附录 A-25)} \\
&= \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t) + \widetilde{\xi}_t;\mathbf{Z}|\widetilde{W}_{t-1}) && \text{(附录 A-26)} \\
&\leq \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t) + \widetilde{\xi}_t;B_t|\widetilde{W}_{t-1}) && \text{(附录 A-27)} \\
&= \sum_{t=1}^{T} I_1(\widetilde{W}_t - \widetilde{W}_{t-1};B_t|\widetilde{W}_{t-1}) = \sum_{t=1}^{T} I_1(\widetilde{W}_t;B_t|\widetilde{W}_{t-1}),
\end{aligned}
$$

其中，式 (附录 A-23) 可通过 Markov 链 $Z \to (X,Y) \to f(X,Y)$ 及定理 2.9 中的性质 6 得到，式 (附录 A-24) 可通过性质 5 得到，式 (附录 A-25) 可通过重复式 (附录 A-24) 以上步骤得到，式 (附录 A-26) 可由 $\widetilde{W}_0$ 与 $\mathbf{Z}$ 的独立性得到，式 (附录 A-27) 可通过 Markov 链 $\mathbf{Z} \to B_t \to -\eta_t g(w,B_t) + \widetilde{\xi}_t | w = \widetilde{W}_{t-1}$ 得到。 ■

证明 (定理 2.16)**:** 应用引理 附录 A.18，有

$$
I_1(\widetilde{W}_T;\mathbf{Z}) \leq \sum_{t=1}^{T} \left(\frac{1}{2} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathrm{Cov}[g(W_{t-1},B_t)] + I \right| + E^{\kappa}_{\widetilde{W}_t|\widetilde{W}_{t-1}} \right),
$$

在引理 附录 A.16 中取 $X = B_t$，$Y = \widetilde{W}_{t-1}$，$\Delta = -\Delta_{t-1}$，$\xi = \widetilde{\xi}_t$ 与 $f(W_{t-1},B_t) = -\eta_t g(W_{t-1},B_t)$，定理得证。 ■

引理 附录 A.19 设 $u,v \in [0,1]$ 分别为核矩阵 A 的最小与最大特征值，且

$$
\mu = \frac{1-un}{v-u} \cdot v^{\alpha} + \frac{vn-1}{v-u} \cdot u^{\alpha}, \qquad \text{(附录 A-28)}
$$

则有

$$
|\log \mu| = \Omega(|\alpha - 1|), \quad |\log \mu| = O(|\alpha - 1| \log n). \qquad \text{(附录 A-29)}
$$

证明: 当 $u = 0$ 时，有 $\mu = v^{\alpha-1}$ 且 $v \in (1/n, 1)$，上述结论显然成立。否则，设 $\kappa = v/u$ 为核矩阵 A 的条件数，则有

$$
\mu = u^{\alpha} \frac{\kappa^{\alpha} - \kappa^{\alpha} un + \kappa un - 1}{\kappa u - u} = u^{\alpha-1} \left(\frac{\kappa(\kappa^{\alpha-1} - 1)(1-un)}{\kappa - 1} + 1 \right).
$$

忽略朴素情形 $\kappa = 1$，则对于任意满足 $\kappa > \gamma + 1$ 的正常数 γ，有

$$
\kappa^{\alpha-1} - 1 \leq \frac{\kappa(\kappa^{\alpha-1} - 1)}{\kappa - 1} \leq \left(1 + \frac{1}{\gamma} \right) (\kappa^{\alpha-1} - 1).
$$

因此，对于任意 $\kappa \in (1,\infty)$，有

$$|\log\mu| = \Theta\left(\left|(\alpha-1)\log u + \log\left((\kappa^{\alpha-1}-1)(1-un)+1\right)\right|\right).$$

当 $u \in (1/2n, 1/n)$ 时，有 $1-un = O(1)$ 且

$$|\log\mu| = \Omega\Big(\big|(\alpha-1)\log u\big|\Big) = \Omega\Big(\big|(\alpha-1)\log n\big|\Big),$$
$$|\log\mu| = O\Big(\big|(\alpha-1)(\log u + \log k)\big|\Big) = O\Big(\big|(\alpha-1)\log v\big|\Big).$$

否则当 $u <= 1/2n$ 时，有 $|\log\mu| = \Theta(|(\alpha-1)\log v|)$。结合 $v \in (1/n, 1)$，可得

$$|\log\mu| = \Omega(|\alpha-1|), \quad |\log\mu| = O(|\alpha-1|\log n). \qquad \blacksquare$$

引理 附录 A.20 对于任意 $\varepsilon_0 \in (0,1)$ 和足够大的 n，若随机算法 $\mathcal{A}$ 能够以 $1-\delta$ 的概率通过 s 次查询给出 $\mathrm{tr}(A)$ 的一个 $(1\pm\varepsilon_0)$ 近似值，则 $\mathcal{A}$ 能够在同等条件下给出 $S_\alpha(A)$ 的 $(1\pm\varepsilon)$ 近似值，其中 $\varepsilon_0 = 1-\min(\mu, 1/\mu)^\varepsilon$。反之，对于 $\varepsilon_0 = \max(n^{\alpha-1}, n^{1-\alpha})^\varepsilon - 1$ 同样成立。

证明: 注意到

$$\mathrm{tr}(A^\alpha) \in \begin{cases} [\mu, n^{1-\alpha}], & \text{若 } \alpha < 1 \\ [n^{1-\alpha}, \mu], & \text{若 } \alpha > 1 \end{cases}, \qquad \text{(附录 A-30)}$$

其中 $\mathrm{tr}(A^\alpha) = \mu$ 对应于其特征值均取 u 或 v 的情形，$\mathrm{tr}(A^\alpha) = n^{1-\alpha}$ 则对应特征值全部取 $\frac{1}{n}$ 的情形。设 Z 为算法 $\mathcal{A}$ 输出的 $\mathrm{tr}(A)$ 的近似值，则以概率 $1-\delta$，有

$$-\varepsilon_0 \cdot \mathrm{tr}(A^\alpha) \le Z - \mathrm{tr}(A^\alpha) \le \varepsilon_0 \cdot \mathrm{tr}(A^\alpha).$$

当 $\alpha < 1$ 时，依据式 (附录 A-30) 有 $1 < \mu \le \mathrm{tr}(A^\alpha)$，故

$$1-\varepsilon_0 = \mu^{-\varepsilon} \ge \mathrm{tr}^{-\varepsilon}(A^\alpha), \qquad 1+\varepsilon_0 < \frac{1}{1-\varepsilon_0} = \mu^\varepsilon \le \mathrm{tr}^\varepsilon(A^\alpha),$$
$$(\mathrm{tr}^{-\varepsilon}(A^\alpha)-1)\mathrm{tr}(A^\alpha) \le Z - \mathrm{tr}(A^\alpha) \le (\mathrm{tr}^{\varepsilon}(A^\alpha)-1)\mathrm{tr}(A^\alpha),$$
$$\mathrm{tr}^{1-\varepsilon}(A^\alpha) \le Z \le \mathrm{tr}^{1+\varepsilon}(A^\alpha), \qquad \mathrm{tr}^{-\varepsilon}(A^\alpha) \le \frac{Z}{\mathrm{tr}(A^\alpha)} \le \mathrm{tr}^{\varepsilon}(A^\alpha).$$

在上述不等式两侧取 log 函数，有

$$\frac{1}{1-\alpha}\left|\log Z - \log\mathrm{tr}(A^\alpha)\right| \le \frac{\varepsilon}{1-\alpha}\left|\log\mathrm{tr}(A^\alpha)\right|, \qquad \left|\widetilde{S}_\alpha(A) - S_\alpha(A)\right| \le \varepsilon \cdot S_\alpha(A),$$

其中 $\widetilde{S}_\alpha(A) = \frac{1}{1-\alpha}\log Z$ 为 $S_\alpha(A)$ 的近似结果。类似地，对于 $\alpha > 1$ 情形有相同结论。反之，设 $\overline{Z}$ 为算法 $\mathcal{A}$ 输出的 $S_\alpha(A)$ 的近似值，则以概率 $1-\delta$，有 $\left|\overline{Z} - S_\alpha(A)\right| \le \varepsilon \cdot S_\alpha(A)$。当 $\alpha < 1$ 时，通过类似推导步骤可得

$$(\mathrm{tr}^{-\varepsilon}(A^\alpha)-1)\mathrm{tr}(A^\alpha)\le \tilde{\mathrm{tr}}(A^\alpha)-\mathrm{tr}(A^\alpha)\le (\mathrm{tr}^{\varepsilon}(A^\alpha)-1)\mathrm{tr}(A^\alpha),$$

其中 $\tilde{\mathrm{tr}}(A^\alpha)=\exp((1-\alpha)\overline{Z})$ 为 $\mathrm{tr}(A^\alpha)$ 的近似结果。由式 (附录 A-30) 知 $n^{1-\alpha}\ge \mathrm{tr}(A^\alpha)$，且

$$\mathrm{tr}^{\varepsilon}(A^\alpha)\le n^{\varepsilon(1-\alpha)}=1+\varepsilon_0,\qquad \mathrm{tr}^{-\varepsilon}(A^\alpha)\ge n^{-\varepsilon(1-\alpha)}=\frac{1}{1+\varepsilon_0}\ge 1-\varepsilon_0.$$

综合上述不等式可得 $-\varepsilon_0\cdot\mathrm{tr}(A^\alpha)\le \tilde{\mathrm{tr}}(A^\alpha)-\mathrm{tr}(A^\alpha)\le \varepsilon_0\cdot\mathrm{tr}(A^\alpha)$。对于 $\alpha>1$ 可得相同结论。定理得证。 ■

引理 附录 A.21 ([188]，定理 1) 若在算法 2-1 中取 $s=O\left(\frac{1}{\varepsilon}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right)$，则对于任意半正定矩阵 $f(A)$，其输出 Z 以概率 $1-\delta$ 满足：$\left|Z-\mathrm{tr}\left(f(A)\right)\right|\le\varepsilon\cdot\mathrm{tr}\left(f(A)\right)$.

证明 (定理 2.17)**:** 结合引理 附录 A.20 与 附录 A.21，取 $s=O\left(\frac{1}{\varepsilon_0}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right)$，则算法 2-2 可以概率 $1-\delta$ 保证其输出 $\widetilde{S}_\alpha(A)$ 满足 $\left|\widetilde{S}_\alpha(A)-S_\alpha(A)\right|\le\varepsilon\cdot S_\alpha(A)$，其中 $\varepsilon_0=1-\min(\mu,1/\mu)^\varepsilon$。进一步结合引理 附录 A.19 可得以下收敛阶：

$$s=O\left(\frac{1}{\varepsilon|\alpha-1|}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right).$$

对于整数 $\alpha\ge 2$ 的情形，有 $|\alpha-1|=\Theta(1)$。 ■

引理 附录 A.22 ([268]，定理 2) 设 $\Gamma(x)$ 为 gamma 函数，且 $R(x,y)=\Gamma(x+y)/\Gamma(x)$，则

$$\begin{aligned}
&R(x,y)\ge x(x+y)^{y-1} && 0\le y\le 1,\\
&R(x,y)\ge x^y && 1\le y\le 2,\\
&R(x,y)\ge x(x+1)^{y-1} && y\ge 2.
\end{aligned}$$

引理 附录 A.23 ([188]，定理 5) 若对于算法 2-1 取 $s=O\left(\frac{1}{\varepsilon}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right)$，则对于任意矩阵 $f(A)$，其输出 Z 以概率 $1-\delta$ 满足：$\left|Z-\mathrm{tr}\left(f(A)\right)\right|\le\varepsilon\cdot\|f(A)\|_*$，其中 $\|\cdot\|_*$ 为核范数。

引理 附录 A.24 ([269]，定理 2.1) 假设函数 f 在由焦点位于 ±1 且长短轴长度之和为 $K>1$ 的椭圆区域内解析且满足 $|f(z)|\le M$。设 p_m 为函数 f 的 m 阶 Chebyshev 级数展开，则对于任意 $m\in\mathbb{Z}^+$，有

$$\max_{\lambda\in[-1,1]}|f(\lambda)-p_m(\lambda)|\le\frac{4M}{(K-1)K^m}.$$

引理 附录 A.25 设 g 为线性映射 $[-1,1]\to[u,v]$，$f(\lambda)=\lambda^\alpha$，$p_m(\lambda)$ 为函数 $f\circ g$ 的 $m=O\left(\sqrt{\frac{v}{u}}\log\left(\frac{v}{u\varepsilon}\right)\right)$ 阶 Chebyshev 级数展开，则有

$$\max_{x\in[-1,1]}|(f\circ g)(x)-p_m(x)|=\max_{\lambda\in[u,v]}|f(\lambda)-q_m(\lambda)|\le\varepsilon u^\alpha,$$

其中 $q_m=p_m\circ g^{-1}$。

证明： 可验证幂函数 $f(z)=z^\alpha$ 在区域 $\mathbb{C}/\{-\infty,0\}$ 内解析。结合线性映射 g，函数 $f\circ g$ 在区域 $\mathbb{C}/\{-\infty,-1-\frac{2u}{v-u}\}$ 内解析。因此，可选择椭圆区域 E_c，其长轴长度为 $1+\frac{2u}{v-u}=1+\beta$，短轴长度为 $\sqrt{(\beta+1)^2-1}=\sqrt{\beta^2+2\beta}$，且焦点位于 ±1。在引理 附录 A.24 中取 $K=1+\beta+\sqrt{\beta^2+2\beta}$ 且 $M=(1+\beta)^\alpha$，且注意到 $\log K=\Theta\left(\sqrt{\beta}\right)$，可得以下上界：

$$m\ge\frac{\log\left(\frac{4M}{(K-1)\varepsilon u^\alpha}\right)}{\log K}=O\left(\sqrt{\frac{v}{u}}\log\left(\frac{v}{u\varepsilon}\right)\right).$$ ■

证明 (定理 2.18)**：** 在引理 附录 A.25 中取 $m=O\left(\sqrt{\frac{v}{u}}\log\left(\frac{v}{u\varepsilon_1}\right)\right)$，其中 $\varepsilon_1=\varepsilon_0/2$，$\varepsilon_0=1-\min(\mu,1/\mu)^\varepsilon$，则有

$$\begin{aligned}&\max_{\lambda\in[u,v]}|f(\lambda)-q_m(\lambda)|\le\varepsilon_1u^\alpha,\\&\left|\mathrm{tr}(q_m(A))-\mathrm{tr}(A^\alpha)\right|\le\sum_{i=1}^n|f(\lambda_i)-q_m(\lambda_i)|\le n\varepsilon_1u^\alpha\le\frac{\varepsilon_0}{2}\cdot\mathrm{tr}(A^\alpha).\end{aligned}$$

另外，注意到

$$\min_{\lambda\in[u,v]}q_m(\lambda)\ge\min_{\lambda\in[u,v]}\lambda^\alpha-\max_{\lambda\in[u,v]}|\lambda^\alpha-q_m(\lambda)|\ge u^\alpha-\frac{\varepsilon_0}{2}u^\alpha\ge0.$$

因此，在 m 足够大时，$q_m(A)$ 为半正定矩阵。

在引理 附录 A.21 中取 $s=O\left(\frac{1}{\varepsilon_2}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right)$，其中 $\varepsilon_2=\frac{\varepsilon_0}{3}$，可得

$$\begin{aligned}\left|Z-\mathrm{tr}(q_m(A))\right|&\le\frac{\varepsilon_0}{3}\mathrm{tr}(q_m(A)),\\\mathrm{tr}(q_m(A))&\le\frac{\varepsilon_0}{2}\mathrm{tr}(A^\alpha)+\mathrm{tr}(A^\alpha)\le\frac{3}{2}\mathrm{tr}(A^\alpha),\end{aligned}$$

其中 Z 为 Hutch++ 算法输出的 $\mathrm{tr}(q_m(A))$ 的估计值。综合以上结果可得

$$\begin{aligned}|Z-\mathrm{tr}(A^\alpha)|&\le\left|Z-\mathrm{tr}(q_m(A))\right|+\left|\mathrm{tr}(q_m(A))-\mathrm{tr}(A^\alpha)\right|\\&\le\frac{\varepsilon_0}{3}\mathrm{tr}(q_m(A))+\frac{\varepsilon_0}{2}\mathrm{tr}(A^\alpha)\le\frac{\varepsilon_0}{2}\mathrm{tr}(A^\alpha)+\frac{\varepsilon_0}{2}\mathrm{tr}(A^\alpha)\le\varepsilon_0\cdot\mathrm{tr}(A^\alpha).\end{aligned}$$

通过引理 附录 A.20 与引理 附录 A.19，可得

$$s=O\left(\frac{1}{\varepsilon|\alpha-1|}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right),\qquad m=O\left(\sqrt{\kappa}\log\left(\frac{\kappa}{\varepsilon|\alpha-1|}\right)\right).$$ ■

证明 (定理 2.19)**：** 当 $u=0$ 时，Chebyshev 级数中的系数 $\widehat{T}_k$ 有以下解析表示：

$$c_k = \frac{2}{\pi}\int_0^{\pi}(q_m)^{\alpha}(cos\theta)\cos(k\theta)\,\mathrm{d}\theta = \frac{2v^{\alpha}\Gamma(\alpha+\frac{1}{2})(\alpha)_k}{\sqrt{\pi}\Gamma(\alpha+1)(\alpha+k)_k}.$$

其中 $(\alpha)_k$ 为递降阶乘：$(\alpha)_k = \alpha\cdot\ldots\cdot(\alpha-k+1)$。则对于核矩阵 A 的任一特征值 λ，有

$$\begin{aligned}
|\lambda^{\alpha} - q_m(\lambda)| &= \left|\sum_{i=m+1}^{\infty} c_i\widehat{T}_i(\lambda)\right| \le \sum_{i=m+1}^{\infty}|c_i| = \sum_{i=m+1}^{\infty}\left|\frac{2v^{\alpha}\Gamma(\alpha+\frac{1}{2})(\alpha)_i}{\sqrt{\pi}\Gamma(\alpha+1)(\alpha+i)_i}\right| && \text{(附录 A-31)}\\
&= \frac{2v^{\alpha}}{\sqrt{\pi}}\sum_{i=m+1}^{\infty}\left|\frac{\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+1)}{\Gamma(\alpha+i+1)\Gamma(\alpha-i+1)}\right|\\
&\le \frac{2v^{\alpha}}{\sqrt{\pi}}\sum_{i=m+1}^{\infty}\left|\frac{\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+1)}{\Gamma(i-\alpha)\Gamma(\alpha-i+1)(i-\alpha)^{2\alpha+1}}\right| && \text{(附录 A-32)}\\
&\le \frac{2v^{\alpha}\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+1)}{\pi^{3/2}}\sum_{i=m+1}^{\infty}\left|\frac{1}{(i-\alpha)^{2\alpha+1}}\right| && \text{(附录 A-33)}\\
&\le \frac{2v^{\alpha}\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+1)}{\pi^{3/2}}\int_m^{\infty}\frac{1}{(x-\alpha)^{2\alpha+1}}\,\mathrm{d}x && \text{(附录 A-34)}\\
&= \frac{2v^{\alpha}\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha+1)}{\pi^{3/2}}\frac{1}{2\alpha(m-\alpha)^{2\alpha}} = \frac{v^{\alpha}\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha)}{\pi^{3/2}(m-\alpha)^{2\alpha}}.
\end{aligned}$$

其中，式 (附录 A-31) 可由以下事实得出：对于任意 $x\in[0,v]$，有 $\widehat{T}_n(x)\in[-1,1]$；式 (附录 A-32) 可通过对 $R(i-\alpha,2\alpha+1)$ 应用引理 附录 A.22 得出；式 (附录 A-33) 可通过 Euler 反射公式得出；式 (附录 A-34) 可由以下事实得出：假设 $m>\alpha$，则对于任意 $n>1$ 与 $k>1$，有 $n^{-k}\le\int_{n-1}^{n}x^{-k}\,\mathrm{d}x$。

设 $\varepsilon_0 = 1-\min(\mu,1/\mu)^{\varepsilon}$ 且 $\varepsilon_1 = \frac{\varepsilon_0}{3}$。通过下式选择 m：

$$\frac{nv^{\alpha}\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha)}{\pi^{3/2}(m-\alpha)^{2\alpha}} \le \frac{\varepsilon_0}{2}\cdot\mathrm{tr}(A^{\alpha}), \qquad m\ge\alpha+\sqrt[2\alpha]{\frac{2nv^{\alpha}\Gamma(\alpha+\frac{1}{2})\Gamma(\alpha)}{\varepsilon_0\pi^{3/2}\min(v^{\alpha-1},n^{1-\alpha})}}.$$

设 $\lambda_1,\cdots,\lambda_n$ 为核矩阵 A 的特征值，则有 $\left|\mathrm{tr}(q_m(A))-\mathrm{tr}(A^{\alpha})\right|\le\frac{\varepsilon_0}{2}\cdot\mathrm{tr}(A^{\alpha})$。在引理 附录 A.23 中取 $s = O\left(\frac{1}{\varepsilon_1}\sqrt{\log\left(\frac{1}{\delta}\right)}+\log\left(\frac{1}{\delta}\right)\right)$，则有

$$\begin{aligned}
\left|Z-\mathrm{tr}(q_m(A))\right| &\le \frac{\varepsilon_0}{3}\|q_m(A)\|_* = \frac{\varepsilon_0}{3}\sum_{i=1}^{n}|q_m(\lambda_i)| \le \frac{\varepsilon_0}{3}\left(\sum_{i=1}^{n}\lambda_i^{\alpha}+\sum_{i=1}^{n}|\lambda_i^{\alpha}-q_m(\lambda_i)|\right)\\
&\le \frac{\varepsilon_0}{3}\left(\mathrm{tr}(A^{\alpha})+\frac{\varepsilon_0}{2}\mathrm{tr}(A^{\alpha})\right) \le \frac{\varepsilon_0}{2}\cdot\mathrm{tr}(A^{\alpha}).
\end{aligned}$$

综合以上结果可得

$$|Z-\mathrm{tr}(A^{\alpha})| \le \left|Z-\mathrm{tr}(p_m(A))\right|+\left|\mathrm{tr}(p_m(A))-\mathrm{tr}(A^{\alpha})\right| \le \frac{\varepsilon_0}{2}\mathrm{tr}(A^{\alpha})+\frac{\varepsilon_0}{2}\mathrm{tr}(A^{\alpha}) = \varepsilon_0\cdot\mathrm{tr}(A^{\alpha}).$$

应用引理 附录 A.20 及引理 附录 A.19，最终可得

$$s = O\left(\frac{1}{\varepsilon|\alpha-1|}\sqrt{\log\left(\frac{1}{\delta}\right)} + \log\left(\frac{1}{\delta}\right)\right), \qquad m = \begin{cases} O\left(\sqrt[2\alpha]{vn}\sqrt[2\alpha]{\frac{1}{\varepsilon|\alpha-1|}}\right) & \alpha < 1 \\ O\left(\sqrt{vn}\sqrt[2\alpha]{\frac{1}{\varepsilon|\alpha-1|}}\right) & \alpha > 1 \end{cases}. \qquad \blacksquare$$

引理 附录 A.26 ([188]，定理 7) 对于任意仅通过矩阵向量乘法 $Ar_1, \cdots, Ar_m$ 访问半正定矩阵 A 的算法，其中 $r_1, \cdots, r_m$ 为算法自适应选择的整数向量，且取值范围为 $[-2^b, \cdots, 2^b]$，至少需要 $s = \Omega\left(\frac{1}{\varepsilon\left(b+\log\left(1/\varepsilon\right)\right)}\right)$ 次矩阵向量乘法查询以获得估计值 Z，使得以至少 $\frac{2}{3}$ 的概率，有 $(1-\varepsilon)\operatorname{tr}(A) \le Z \le (1+\varepsilon)\operatorname{tr}(A)$。

证明 (定理 2.20)**:** 结合引理 附录 A.26 与引理 附录 A.20，可知以概率 $\frac{2}{3}$ 计算 $S_\alpha(A)$ 的 $1\pm\varepsilon$ 近似所需的随机向量数量至少为 $s = \Omega\left(\frac{1}{\varepsilon_0\left(b+\log\left(1/\varepsilon_0\right)\right)}\right)$，其中 $\varepsilon_0 = n^{\varepsilon|\alpha-1|} - 1 \le \varepsilon|1-\alpha|\log n$。在有限精度计算模型下，$b$ 可视为常数，故可得

$$s = \Omega\left(\frac{1}{\varepsilon|\alpha-1|\log n \log\left(\frac{1}{\varepsilon|\alpha-1|\log n}\right)}\right). \qquad \blacksquare$$

引理 附录 A.27 ([270]，[271]) 设 $\|\cdot\|$ 为函数的 L_∞ 范数，$E_m(f) = \min_{p\in\mathbb{P}_m}\|f-p\|$ 表示在有限区间 $[-1,1]$ 上对于给定函数 $f(x)$ 的最优一致逼近误差。则当 $m\to\infty$ 时，有

$$E_m\left((\gamma-x)^{-t}\right) \sim \frac{m^{t-1}}{|\Gamma(t)|}\frac{\left(\gamma-\sqrt{\gamma^2-1}\right)^m}{\left(\sqrt{\gamma^2-1}\right)^{1+t}},$$

其中 $t,\gamma\in\mathbb{R}$ 且 $\gamma>1$。

引理 附录 A.28 存在正值递减函数 $\varepsilon_0:\mathbb{R}^+\to\mathbb{R}^+$，使得对于任意 $0<u<v<1$ 和 $\varepsilon\in(0,\varepsilon_0(v/u))$，任意多项式 $q_m(\lambda)$ 至少需要 $m=\Omega\left(\sqrt{\frac{v}{u}}\log\left(\frac{u}{v\varepsilon}\right)\right)$ 阶以实现对于任意满足 $\sum_{i=1}^n\lambda_i\in[b,b+v)$ 实数序列 $\lambda_1,\cdots,\lambda_n\in[u,v]$，有 $\left|\sum_{i=1}^n(f(\lambda_i)-q_m(\lambda_i))\right|\le\varepsilon$，其中 $b\ge v$ 为常数。

证明: 在相同假设条件下，列出以下的多项式近似复杂度问题。以下将证明以下每个问题的复杂度均可顺序归约到下一个问题：

问题 附录 A.29 多项式 q_m 实现以下条件所需的最小阶数：对于任意满足 $\sum_{i=1}^n\lambda_i\in[b,b+v)$ 的序列 $\lambda_1,\cdots,\lambda_n\in[u,v]$，均有 $\left|\sum_{i=1}^n(f(\lambda_i)-q_m(\lambda_i))\right|\le\varepsilon$。

问题 附录 A.30 多项式 q_m 实现以下条件所需的最小阶数：对于任意满足 $\sum_{i=1}^n\lambda_i=b$ 的序列 $\lambda_1,\cdots,\lambda_n\in[u,v]$，均有 $\sum_{i=1}^n|f(\lambda_i)-q_m(\lambda_i)|\le\varepsilon$。

问题 附录 A.31 多项式 q_m 实现以下条件所需的最小阶数：对于任意 $\lambda\in[u,v]$，均有

$|f(\lambda)-q_m(\lambda)|\varphi(\lambda)\le\varepsilon$，其中 $\varphi(\lambda)=\lfloor\min(\frac{nv-b}{\lambda-u},\frac{b-nu}{v-\lambda})\rfloor$。

问题 附录 A.32 多项式 q_m 实现以下条件所需的最小阶数：对于任意 $\lambda\in[u,v]$，均有 $|f(\lambda)-q_m(\lambda)|\le\varepsilon$。

问题 附录 A.33 多项式 q_m 实现以下条件所需的最小阶数：对于任意 $\lambda\in[u,u+2]$，均有 $|f(\lambda)-q_m(\lambda)|\le\varepsilon$。

对于问题 附录 A.33，由于以下函数关于 y 轴的对称性，近似函数 $(f\circ g)(x)=(x+u+1)^{\alpha}$ 等价于近似 $(\gamma-x)^{-t}$，其中 $\gamma=u+1$、$t=-\alpha$。设 $\varepsilon=E_m(f\circ g)$，基于以下 gamma 函数性质：$|\Gamma(-\alpha)|=\frac{\pi}{|\sin\pi\alpha|\Gamma(\alpha+1)}$，进一步应用引理 附录 A.27，当 m 足够大时，有

$$\lim_{m\to\infty}\frac{\left(1+u-\sqrt{u^2+2u}\right)^m}{m^{\alpha+1}}=\frac{\varepsilon|\Gamma(-\alpha)|}{\left(\sqrt{u^2+2u}\right)^{\alpha-1}}=\Theta\left(\frac{\varepsilon}{(\sqrt{u})^{\alpha-1}|sin\pi\alpha|}\right).$$

因此，对于任意 $u\in(0,1)$，存在 $\varepsilon_0\in(0,1)$，使得当 $\varepsilon<\varepsilon_0$ 时，有

$$m=\Omega\left(\frac{1}{\sqrt{u}}\log\left(\frac{u|\sin\pi\alpha|}{\varepsilon}\right)\right).$$

当 $\alpha\notin\mathbb{N}$ 时，有 $|\sin\pi\alpha|=\Theta(1)$。

问题 附录 A.32 → 附录 A.33：近似 $f_{[u,u+2]}(\lambda)$ 可通过近似 $\left(\frac{u+2}{v}\right)^{\alpha}f_{[\frac{uv}{u+2},v]}\left(\frac{v}{u+2}\lambda\right)$ 实现，因此近似函数 $f_{[\frac{uv}{u+2},v]}$ 的多项式阶数下界为 $m=\Omega\left(\frac{1}{\sqrt{u}}\log\left(\frac{u}{v\varepsilon}\right)\right)$。其等价于通过以下阶数近似函数 $f_{[u,v]}$：$m=\Omega\left(\sqrt{\frac{v}{u}}\log\left(\frac{u}{v\varepsilon}\right)\right)$。

问题 附录 A.31 → 附录 A.32：不妨设 $u<\frac{b-v}{n-1}$ 且 $v>\frac{b-u}{n-1}$，否则将有 $\min_i\lambda_i>u$ 或 $\max_i\lambda_i<v$，故可进一步缩减 $[u,v]$ 范围以满足此假设。则对于任意 $\lambda\in[u,v]$，有

$$\varphi(\lambda)=\left\lfloor\min\left(\frac{nv-b}{v-\lambda},\frac{b-nu}{\lambda-u}\right)\right\rfloor\ge\left\lfloor\min\left(\frac{nv-b}{v-u},\frac{b-nu}{v-u}\right)\right\rfloor\ge1.$$

问题 附录 A.30 → 附录 A.31：对于任意 $\lambda\in[u,v]$，构造序列：$\lambda_1,\cdots,\lambda_n$，其中 $\lambda_1,\cdots,\lambda_{n_\lambda}=\lambda$，$\lambda_{n_\lambda+1},\cdots,\lambda_n=\frac{b-\lambda n_\lambda}{n-n_\lambda}$，$n_\lambda=\lfloor\min(\frac{nv-b}{\lambda-u},\frac{b-nu}{v-\lambda})\rfloor$。则此序列满足 $\sum_{i=1}^n\lambda_i=b$，

$$\sum_{i=1}^n|f(\lambda_i)-q_m(\lambda_i)|\le\varepsilon,\quad n_\lambda|f(\lambda)-q_m(\lambda)|\le\varepsilon,\quad\varphi(\lambda)|f(\lambda)-q_m(\lambda)|\le\varepsilon.$$

问题 附录 A.29 → 附录 A.30：设 q_m 为问题 附录 A.29 的解，实现了 $\varepsilon/2$ 近似精度。朴素情形下，对于任意 $\lambda\in[u,v]$ 有 $f(\lambda)\le q_m(\lambda)$ 或 $f(\lambda)\ge q_m(\lambda)$，则 $\sum_{i=1}^n|f(\lambda_i)-q_m(\lambda_i)|=\left|\sum_{i=1}^n(f(\lambda_i)-q_m(\lambda_i))\right|$。

否则，由于函数 $f(\lambda)$ 与 $q_m(\lambda)$ 的连续性，将存在 $\rho\in(u,v)$ 使得 $f(\rho)=q_m(\rho)$。给定

任意 $\lambda_1,\cdots,\lambda_n$，存在对于序列 λ_i 重排后的一种划分 n_ρ，使得对于任意 $i\in[1,n_\rho]$，有 $f(\lambda_i)\le q_m(\lambda_i)$，且对于任意 $i\in[n_\rho+1,n]$，有 $f(\lambda_i)>q_m(\lambda_i)$。

$$\sum_{i=1}^{n_\rho}|f(\lambda_i)-q_m(\lambda_i)|=\left|\sum_{i=1}^{n_\rho}(f(\lambda_i)-q_m(\lambda_i))\right|,$$
$$\sum_{i=n_\rho+1}^{n}|f(\lambda_i)-q_m(\lambda_i)|=\left|\sum_{i=n_\rho+1}^{n}(f(\lambda_i)-q_m(\lambda_i))\right|.$$

构造序列：$\lambda_1^1,\cdots,\lambda_{n_1}^1$ 与 $\lambda_1^2,\cdots,\lambda_{n_2}^2$，其中

$$\lambda_i^1=\begin{cases}\lambda_i & i\le n_\rho\\ \rho & i>n_\rho\end{cases},\quad n_1=n_\rho+\left\lceil\sum_{i=n_\rho+1}^{n}\lambda_i/\rho\right\rceil.$$
$$\lambda_i^2=\begin{cases}\lambda_{i+n_\rho} & i\le n-n_\rho\\ \rho & i>n-n_\rho\end{cases},\quad n_2=n-n_\rho+\left\lceil\sum_{i=1}^{n_\rho}\lambda_i/\rho\right\rceil.$$

设 $b=\sum_{i=1}^{n}\lambda_i$，则 $\sum_{i=1}^{n_1}\lambda_i^1\in[b,b+v)$ 且 $\sum_{i=1}^{n_2}\lambda_i^2\in[b,b+v)$。因此

$$\begin{aligned}\sum_{i=1}^{n}|f(\lambda_i)-q_m(\lambda_i)|&=\sum_{i=1}^{n_\rho}|f(\lambda_i)-q_m(\lambda_i)|+\sum_{i=n_\rho+1}^{n}|f(\lambda_i)-q_m(\lambda_i)|\\&=\left|\sum_{i=1}^{n_\rho}(f(\lambda_i)-q_m(\lambda_i))\right|+\left|\sum_{i=n_\rho+1}^{n}(f(\lambda_i)-q_m(\lambda_i))\right|\\&=\left|\sum_{i=1}^{n_1}(f(\lambda_i^1)-q_m(\lambda_i^1))\right|+\left|\sum_{i=1}^{n_2}(f(\lambda_i^2)-q_m(\lambda_i^2))\right|\le\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.\end{aligned}$$

综合以上复杂度归约过程，可得问题 附录 A.29 的复杂度下界 $m=\Omega\left(\sqrt{\frac{v}{u}}\log\left(\frac{u}{v\varepsilon}\right)\right)$。 ■

证明 (定理 2.21)**:** 设 $Z=\mathrm{tr}(q_m(A))$，则有 $|Z-\mathrm{tr}(A^\alpha)|=\left|\sum_{i=1}^{n}(q_m(\lambda_i)-\lambda_i^\alpha)\right|$。设 $\varepsilon_0=n^{\varepsilon|\alpha-1|}-1$，通过引理 附录 A.28，可得 $m=\Omega\left(\sqrt{\frac{v}{u}}\log\left(\frac{u}{v\varepsilon_0}\right)\right)$ 为实现 $|Z-\mathrm{tr}(A^\alpha)|\le\varepsilon_0$ 的复杂度下界，其中 $b=1$。结合引理 附录 A.20，近似矩阵多项式的阶数下界为：

$$m=\Omega\left(\sqrt{\frac{v}{u}}\log\left(\frac{u}{v\varepsilon_0}\right)\right)=\Omega\left(\sqrt{\frac{v}{u}}\log\left(\frac{u}{v\varepsilon|\alpha-1|\log n}\right)\right).$$ ■

引理 附录 A.34 ([272]，[273]) 当 $m\to\infty$ 时，有 $E_m(x^\alpha)\sim\delta(\alpha)m^{-2\alpha}$，其中 $\alpha\in\mathbb{R}^+$，$\delta(\alpha)$ 是仅依赖于 α 的正数，且在 $\alpha\notin\mathbb{N}$ 时满足 $\delta(\alpha)>0$。

引理 附录 A.35 对任意 $v>0$ 和足够小的 ε，任意多项式 $q_m(\lambda)$ 至少需要 $m=\Omega\left(\sqrt[2\alpha]{\frac{1}{\varepsilon}}\right)$

阶以实现对于任意满足 $\sum_{i=1}^{n}\lambda_i\in[b,b+v)$ 的实数序列 $\lambda_1,\cdots,\lambda_n\in[0,v]$，其中 $b\geq v$ 为常数，有 $\left|\sum_{i=1}^{n}(f(\lambda_i)-q_m(\lambda_i))\right|\leq\varepsilon$。

证明： 本引理可通过与引理 附录 A.28 相似的归约过程证明。 ■

证明 (定理 2.22)**：** 设 $Z=\mathrm{tr}(q_m(A))$ 为近似矩阵多项式的迹，$\varepsilon_0=n^{\varepsilon|\alpha-1|}-1$。根据引理 附录 A.35 可知 $m=\Omega\left(\sqrt[2\alpha]{\frac{1}{\varepsilon_0}}\right)$ 为实现 $|Z-\mathrm{tr}(A^{\alpha})|\leq\varepsilon_0$ 的复杂度下界，其中 $b=1$。结合引理 附录 A.20，可得下界 $m=\Omega\left(\sqrt[2\alpha]{\frac{1}{\varepsilon_0}}\right)=\Omega\left(\sqrt[2\alpha]{\frac{1}{\varepsilon|\alpha-1|\log n}}\right)$。 ■

A.3 第 3 章定理补充证明

引理 附录 A.36 ([274], 引理 3) 设向量 $X=(X_1,\cdots,X_n)$ 服从参数为 m 与 $p=(p_1,\cdots,p_n)$ 的多项式分布，$\bar{a}_1,\cdots,\bar{a}_n\geq 0$ 为满足 $\sum_{i=1}^{n}\bar{a}_ip_i\neq 0$ 的常数，则对于任意 $\varepsilon>0$，有

$$\mathbb{P}\left(\sum_{i=1}^{n}\bar{a}_i\left(p_i-\frac{X_i}{m}\right)>\varepsilon\right)\leq\exp\left(-\frac{m\varepsilon^2}{\beta}\right),$$

其中 $\beta=2\sum_{i=1}^{n}\bar{a}_i^2p_i$。

通过 $Z=(\theta(Y,V),Y)$ 定义数据生成分布，其中 Y 为随机生成的标签，θ 为隐式确定型映射且 $V=\{V_i\}_{i=1}^{m}\in\mathcal{V}\subset\mathbb{R}^m$ 为独立同分布扰动变量。对于任意 $y\in\mathcal{Y}$，定义 $c_{i,y}^{w}$ 为模型对于扰动变量 V_i 的敏感度：

$$\begin{aligned}c_{i,y}^{w}=\sup_{v_1,\cdots,v_i,\hat{v}_i,\cdots,v_m}\Big|&\log(p_l\circ\ell)\Big(w,\theta_y(v_1,\cdots,v_i,\cdots,v_m),y\Big)\\&-\log(p_l\circ\ell)\Big(w,\theta_y(v_1,\cdots,\hat{v}_i,\cdots,v_m),y\Big)\Big|,\end{aligned}$$

其中 $p_l(l)=\mathbb{P}(L^w=l)$ 且 $\theta_y(v)=\theta(y,v)$。继而定义模型 w 的全局敏感度为：

$$c_y^{w}=\sup_{i\in[1,m]}c_{i,y}^{w},\quad\text{and}\quad c^{w}=\mathbb{E}_Y[c_Y^{w}].$$

给定假设 $w\in\mathcal{W}$，使用 $\mathcal{L}^w$ 代表损失函数所有可能取值的集合：

$$\mathcal{L}^{w}=\Big\{\ell(w,\theta_y(v),y):v\in\mathcal{V},y\in\mathcal{Y}\Big\}.$$

给定任意 $\gamma>0$，下式定义了损失空间 $\mathcal{L}^w$ 的典型子集：

$$\mathcal{L}_{\gamma}^{w}=\left\{l\in\mathcal{L}^{w}:-\log p_l(l)-H(L^w)\leq c^{w}\sqrt{\frac{m\log\left(\sqrt{n}/\gamma\right)}{2}}\right\}.\qquad\text{(附录 A-35)}$$

注意此处 $\mathcal{L}_{\gamma}^{w}$ 的取值可通过给定假设 $w\in\mathcal{W}$ 而完全确定。

引理 附录 A.37 给定任意 $\gamma>0$，有 $\mathbb{P}\left(L^w \notin \mathcal{L}_\gamma^w\right) \leq \frac{\gamma}{\sqrt{n}}$ 且

$$\left|\mathcal{L}_\gamma^w\right| \leq \exp\left(H(L^w)+c^w\sqrt{\frac{m\log(\sqrt{n}/\gamma)}{2}}\right).$$

证明: 考虑以下函数：

$$f(y,v)=-\log p_l(h_y(v)), \qquad h_y(v)=\ell(w,\theta_y(v)).$$

设 $p_y(y)=\mathbb{P}(Y=y)$，$p_v(v)=\mathbb{P}(V=v)$ 且 $h_y^{-1}(l)=\{v\in\mathcal{V}: h_y(v)=l\}$，则有

$$\begin{aligned}
\mathbb{E}_{Y,V}[f(Y,V)] &= -\sum_{y\in\mathcal{Y}} p_y(y)\sum_{v\in\mathcal{V}} p_v(v)\log p_l(h_y(v)) \\
&= -\sum_{y\in\mathcal{Y}} p_y(y)\sum_{l\in\mathcal{L}^w}\sum_{v\in h_y^{-1}(l)} p_v(v)\log p_l(h_y(v)) \\
&= -\sum_{l\in\mathcal{L}^w}\left(\sum_{y\in\mathcal{Y}} p_y(y)\sum_{v\in h_y^{-1}(l)} p_v(v)\right)\log p_l(l) \\
&= -\sum_{l\in\mathcal{L}^w} p_l(l)\log p_l(l) = H(L^w).
\end{aligned}$$

因此，通过对 $f(V)=-\log p_l(L^w)$ 应用 McDiarmid 不等式，可得

$$\mathbb{P}(-\log p_l(L^w)-H(L^w)\geq\varepsilon)\leq\exp\left(-\frac{2\varepsilon^2}{m(c^w)^2}\right). \qquad \text{(附录 A-36)}$$

令 δ 等于式 (附录 A-36) 右侧，则有 $\varepsilon = c^w\sqrt{\frac{m\log(1/\delta)}{2}}$。结合式 (附录 A-35)，可选择 $\delta=\gamma/\sqrt{n}$ 及 $\mathbb{P}\left(L^w\notin\mathcal{L}_\gamma^w\right)\leq\delta=\frac{\gamma}{\sqrt{n}}$。

现在考虑该典型子集的基数。对于任意 $l\in\mathcal{L}_\gamma^w$，解不等式 $-\log p_l(l)-H(L^w)\leq\varepsilon$ 可得 $\exp(-H(L^w)-\varepsilon)\leq p_l(l)$，进而有

$$1\geq\mathbb{P}\left(L^w\in\mathcal{L}_\gamma^w\right)=\sum_{l\in\mathcal{L}_\gamma^w} p_l(l)\geq\sum_{l\in\mathcal{L}_\gamma^w}\exp(-H(L^w)-\varepsilon)=\left|\mathcal{L}_\gamma^w\right|\exp(-H(L^w)-\varepsilon).$$

结合以上结果，可得 $\left|\mathcal{L}_\gamma^w\right|\leq\exp\left(H(L^w)+c^w\sqrt{\frac{m\log(\sqrt{n}/\gamma)}{2}}\right)$，证毕。 ■

简便起见，定义 $t=\left|\mathcal{L}_\gamma^w\right|$，$\mathcal{L}_\gamma^w=\{a_1,\cdots,a_t\}$ 为典型子集中的元素，且有

$$\mathcal{I}=\left\{i\in[1,n]: L_i^w\notin\mathcal{L}_\gamma^w\right\}, \qquad \mathcal{I}_k=\left\{i\in[1,n]: L_i^w=a_k\right\}.$$

注意此处 $\mathcal{I}$ 与 $\mathcal{I}_k$ 为随机变量，其随机性来源于数据集 $\mathbf{Z}$。

引理 附录 A.38 泛化误差可分解为 $\overline{\text{gen}}(w,\mathbf{Z})=A(w,\mathbf{Z})+B(w,\mathbf{Z})+C(w,\mathbf{Z})$，其中

$$A(w,\mathbf{Z}) = \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\left(\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} L_i^w\right),$$
$$B(w,\mathbf{Z}) = \frac{1}{|\mathcal{I}|}\left(\mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big) - \frac{|\mathcal{I}|}{n}\right)\sum_{i\in\mathcal{I}} L_i^w,$$
$$C(w,\mathbf{Z}) = \sum_{k=1}^{t}\left(\mathbb{P}(L^w = a_k) - \frac{|\mathcal{I}_k|}{n}\right)a_k.$$

证明: 注意到 $\mathcal{I}\cup\mathcal{I}_1\cup\cdots\cup\mathcal{I}_t = n$，则总体风险可分解为

$$\begin{aligned}\mathbb{E}_{L^w}[L^w] &= \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] + \sum_{k=1}^{t}\mathbb{P}(L^w = a_k)\mathbb{E}_{L^w}[L^w | L^w = a_k] \\ &= \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] + \sum_{k=1}^{t}\mathbb{P}(L^w = a_k)a_k. \qquad \text{(附录 A-37)}\end{aligned}$$

类似地，可将经验风险分解为

$$\frac{1}{n}\sum_{i=1}^{n} L_i^w = \frac{1}{n}\left(\sum_{i\in\mathcal{I}} L_i^w + \sum_{k=1}^{t}\sum_{i\in\mathcal{I}_k} L_i^w\right) = \frac{1}{n}\sum_{i\in\mathcal{I}} L_i^w + \sum_{k=1}^{t}\frac{1}{n}\sum_{i\in\mathcal{I}_k} a_k = \frac{1}{n}\sum_{i\in\mathcal{I}} L_i^w + \sum_{k=1}^{t}\frac{|\mathcal{I}_k|}{n}a_k. \qquad \text{(附录 A-38)}$$

将式 (附录 A-37) 与式 (附录 A-38) 代入 $\overline{\text{gen}}(w,\mathbf{Z})$，则有

$$\begin{aligned}\overline{\text{gen}}(w,\mathbf{Z}) &= L(w) - L_{\mathbf{Z}}(w) = \mathbb{E}_{L^w}[L^w] - \frac{1}{n}\sum_{i=1}^{n} L_i^w \\ &= \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] - \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} L_i^w \\ &\quad + \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} L_i^w - \frac{1}{n}\sum_{i\in\mathcal{I}} L_i^w + \sum_{k=1}^{t}\mathbb{P}(L^w = a_k)a_k - \sum_{k=1}^{t}\frac{|\mathcal{I}_k|}{n}a_k \\ &= \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\left(\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} L_i^w\right) \\ &\quad + \frac{1}{|\mathcal{I}|}\left(\mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big) - \frac{|\mathcal{I}|}{n}\right)\sum_{i\in\mathcal{I}} L_i^w + \sum_{k=1}^{t}\left(\mathbb{P}(L^w = a_k) - \frac{|\mathcal{I}_k|}{n}\right)a_k. \qquad \blacksquare\end{aligned}$$

引理 附录 A.39 给定任意 $\gamma > 0$，有 $A(w,\mathbf{Z}) \le \frac{\gamma b^w}{\sqrt{n}}$。

证明: 根据引理 附录 A.37，有 $\mathbb{P}\Big(L_w \notin \mathcal{L}_\gamma^w\Big) \le \frac{\gamma}{\sqrt{n}}$。由于 $L_i^w \ge 0$，故有

$$\begin{aligned}A(w,\mathbf{Z}) &= \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\left(\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} L_i^w\right) \\ &\le \mathbb{P}\Big(L^w \notin \mathcal{L}_\gamma^w\Big)\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] \le \frac{\gamma}{\sqrt{n}}\mathbb{E}_{L^w}\Big[L^w | L^w \notin \mathcal{L}_\gamma^w\Big] \le \frac{\gamma b^w}{\sqrt{n}}. \qquad \blacksquare\end{aligned}$$

引理 附录 A.40 给定任意 $\gamma > 0$ 与 $\delta > 0$，则以置信度 $1-\delta$，有

$$B(w,\mathbf{Z}) \le \frac{\sqrt{\mathbb{P}\left(L^w \notin \mathcal{L}_\gamma^w\right)}\sum_{i\in\mathcal{I}} L_i^w}{|\mathcal{I}|}\sqrt{\frac{2\log(2/\delta)}{n}},$$
$$C(w,\mathbf{Z}) \le 2b^w\sqrt{\frac{2(H(L^w)+C_4^w)+2\log(2/\delta)}{n}},$$

其中 $C_4^w = c^w\sqrt{\frac{m\log(\sqrt{n}/\gamma)}{2}}$。

证明: 设 $q_k = \mathbb{P}(L^w = a_k), q = \mathbb{P}\left(L^w \notin \mathcal{L}_\gamma^w\right)$ 且 $C_k(w,\mathbf{Z}) = \sum_{i=1}^t\left(q_i - \frac{|\mathcal{I}_i|}{n}\right)a_i - \left(q_k - \frac{|\mathcal{I}_k|}{n}\right)a_k$。应用引理 附录 A.36，其中

$$n = t+1, \qquad X = (|\mathcal{I}_1|,\cdots,|\mathcal{I}_t|,|\mathcal{I}|), \qquad p = (q_1,\cdots,q_t,q),$$
$$m = n, \qquad \bar{a}_k = 0,\ \bar{a}_{t+1} = 0,\ \text{and}\ \bar{a}_i = a_i\ \text{for any}\ i \ne k.$$

若存在 $i \in [1,t]\setminus k$ 使得 $q_i a_i > 0$，则有 $\sum_{i=1}^t \bar{a}_i q_i + \bar{a}_{t+1}q \ne 0$ 且引理 附录 A.36 的前置条件得以满足。则对于任意 $\varepsilon > 0$ 与 $k \in [1,t]$，有

$$\mathbb{P}(C_k(w,\mathbf{Z}) > \varepsilon) \le \exp\left(-\frac{n\varepsilon^2}{2\left(\sum_{i=1}^t q_i a_i^2 - q_k a_k^2\right)}\right), \tag{附录 A-39}$$

类似地，设 $\bar{a}_{t+1} = 1$ 且 $\bar{a}_i = 0$，可得

$$\mathbb{P}\left(q - \frac{|\mathcal{I}|}{n} > \varepsilon\right) \le \exp\left(-\frac{n\varepsilon^2}{2q}\right). \tag{附录 A-40}$$

分别令 δ 等于式 (附录 A-39) 与式 (附录 A-40) 右侧，则对于任意 $k \in [1,t]$，有

$$\mathbb{P}\left(C_k(w,\mathbf{Z}) > \sqrt{\sum_{i=1}^t q_i a_i^2 - q_k a_k^2}\sqrt{\frac{2\log(1/\delta)}{n}}\right) \le \delta, \tag{附录 A-41}$$
$$\mathbb{P}\left(q - \frac{|\mathcal{I}|}{n} > \sqrt{\frac{2q\log(1/\delta)}{n}}\right) \le \delta. \tag{附录 A-42}$$

否则若对于任意 $i \ne k$ 均有 $q_i a_i = 0$ 或 $q = 0$，则有 $C_k(w,\mathbf{Z}) = 0$ 或 $q - |\mathcal{I}|/n = 0$，故式 (附录 A-41) 与式 (附录 A-42) 自然成立。因此，上式对任意 $(q_1a_1,\cdots,q_ta_t,q)$ 均成立。

将式 (附录 A-42) 代入 $B(w,\mathbf{Z})$，则对于任意 $\delta > 0$，以置信度 $1-\delta$，有

$$B(w,\mathbf{Z}) = \frac{1}{|\mathcal{I}|}\left(\mathbb{P}\left(L^w \notin \mathcal{L}_\gamma^w\right) - \frac{|\mathcal{I}|}{n}\right)\sum_{i\in\mathcal{I}} L_i^w \le \frac{\sqrt{\mathbb{P}\left(L^w \notin \mathcal{L}_\gamma^w\right)}\sum_{i\in\mathcal{I}} L_i^w}{|\mathcal{I}|}\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{附录 A-43}$$

类似地，根据式 (附录 A-41)，对于任意 $\delta > 0$ 与 $k \in [1,t]$，以置信度 $1-\delta$ 有

$$
\begin{aligned}
C_k(w,\mathbf{Z}) &\le \sqrt{\sum_{i=1}^{t}\mathbb{P}(L^w=a_i)a_i^2-\mathbb{P}(L^w=a_k)a_k^2}\sqrt{\frac{2\log(1/\delta)}{n}}\\
&\le b^w\sqrt{\sum_{i=1}^{t}\mathbb{P}(L^w=a_i)-\mathbb{P}(L^w=a_k)}\sqrt{\frac{2\log(1/\delta)}{n}}\\
&= b^w\sqrt{\mathbb{P}\Big(L^w\in\mathcal{L}_\gamma^w\bigcap L^w\neq a_k\Big)}\sqrt{\frac{2\log(1/\delta)}{n}} \le b^w\sqrt{\frac{2\log(1/\delta)}{n}}.
\end{aligned}
$$

对 $k\in[1,t]$ 取联合上界，则对于任意 $\delta>0$，以置信度 $1-\delta$，下列不等式同时成立

$$
C_1(w,\mathbf{Z})\le b^w\sqrt{\frac{2\log(t/\delta)}{n}},\quad\cdots,\quad C_t(w,\mathbf{Z})\le b^w\sqrt{\frac{2\log(t/\delta)}{n}}. \tag{附录 A-44}
$$

将上式代入 $C(w,\mathbf{Z})$，则对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$
\begin{aligned}
C(w,\mathbf{Z}) &= \sum_{k=1}^{t}\left(\mathbb{P}(L^w=a_k)-\frac{|\mathcal{I}_k|}{n}\right)a_k=\frac{1}{t-1}\sum_{k=1}^{t}C_k(w,\mathbf{Z})\\
&\le \frac{1}{t-1}\sum_{k=1}^{t}b^w\sqrt{\frac{2\log(t/\delta)}{n}}=\frac{t}{t-1}b^w\sqrt{\frac{2\log(t/\delta)}{n}}.
\end{aligned}
$$

对于 $t=1$ 的极限情形，可通过类似方法证明对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$
C(w,\mathbf{Z})=\left(\mathbb{P}(L^w=a_1)-\frac{|\mathcal{I}_1|}{n}\right)a_1\le a_1\sqrt{\mathbb{P}(L^w=a_1)}\sqrt{\frac{2\log(1/\delta)}{n}}\le b^w\sqrt{\frac{2\log(t/\delta)}{n}}.
$$

因此，对于任意 $t\ge1$，有 $C(w,\mathbf{Z})\le 2b^w\sqrt{\frac{2\log(t/\delta)}{n}}$。进一步应用引理 附录 A.37，可得

$$
C(w,\mathbf{Z})\le 2b^w\sqrt{\frac{2\log(t)+2\log(1/\delta)}{n}}\le 2b^w\sqrt{\frac{2\left(H(L^w)+c^w\sqrt{\frac{m\log(\sqrt{n}/\gamma)}{2}}\right)+2\log(1/\delta)}{n}}. \tag{附录 A-45}
$$

取式 (附录 A-43) 与式 (附录 A-45) 的联合不等式，则引理得证。 ■

证明 (定理 3.1)**:** 根据引理 附录 A.39 与引理 附录 A.37，对于任意 $\gamma>0$ 有 $A(w,\mathbf{Z})\le\frac{\gamma b^w}{\sqrt{n}}$ 且 $\mathbb{P}\Big(L^w\notin\mathcal{L}_\gamma^w\Big)\le\frac{\gamma}{\sqrt{n}}$。应用引理 附录 A.40，对任意 $\gamma>0$ 与 $\delta>0$，以置信度 $1-\delta$ 有

$$
\begin{aligned}
B(w,\mathbf{Z}) &\le \frac{\sqrt{\mathbb{P}\Big(L^w\notin\mathcal{L}_\gamma^w\Big)}\sum_{i\in\mathcal{I}}L_i^w}{|\mathcal{I}|}\sqrt{\frac{2\log(2/\delta)}{n}}\\
&\le \frac{\sqrt{\gamma}}{n^{1/4}}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}B^{w,\mathbf{Z}}\sqrt{\frac{2\log(2/\delta)}{n}}=B^{w,\mathbf{Z}}\frac{\sqrt{\gamma}}{n^{1/4}}\sqrt{\frac{2\log(2/\delta)}{n}}, \qquad \text{(附录 A-46)}\\
C(w,\mathbf{Z}) &\le 2b^w\sqrt{\frac{2(H(L^w)+C_4^w)+2\log(2/\delta)}{n}}=C_1^w\sqrt{\frac{H(L^w)+C_2^w}{n}}. \qquad \text{(附录 A-47)}
\end{aligned}
$$

将上式代入引理 附录 A.38，可得

$$\begin{aligned}\overline{\mathrm{gen}}(w,\mathbf{Z}) &\le \frac{\gamma b^w}{\sqrt{n}} + B^{w,\mathbf{Z}}\frac{\sqrt{\gamma}}{n^{1/4}}\sqrt{\frac{2\log(2/\delta)}{n}} + C_1^w\sqrt{\frac{H(L^w)+C_2^w}{n}} \\ &= C_1^w\sqrt{\frac{H(L^w)+C_2^w}{n}} + \frac{1}{\sqrt{n}}\left(\gamma b^w + B^{w,\mathbf{Z}}\frac{\sqrt{\gamma}}{n^{1/4}}\sqrt{2\log(2/\delta)}\right) \\ &= C_1^w\sqrt{\frac{H(L^w)+C_2^w}{n}} + \frac{C_3^w}{\sqrt{n}}. \end{aligned}$$ ■

证明 (定理 3.2)**:** 该定理可通过与定理 3.1 相似的证明过程得到。 ■

给定离散型随机变量 $R \in \mathcal{R}$，设 $p_r(r) = \mathbb{P}(R = r)$。对于任意 $\lambda > 0$，定义 $C_\lambda = \frac{1}{e^{\lambda H(R)}}\sum_{r\in\mathcal{R}} p_r^{1-\lambda}(r)$。对于任意 $\varepsilon > 0$，定义其典型子集为

$$\mathcal{R}_\varepsilon = \{r \in \mathcal{R} : -\log p_r(r) - H(R) \le \varepsilon\}.$$

引理 附录 A.41 对于任意 $\lambda > 0$，取 $\varepsilon = \frac{1}{\lambda}\log(C_\lambda/\delta)$，则有 $\mathbb{P}(R \notin \mathcal{R}_\varepsilon) \le \delta$ 且

$$|\mathcal{R}_\varepsilon| \le \exp\left(H_{1-\lambda}(R) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right).$$

证明: 根据以上典型子集定义，应用 Markov 不等式，有

$$\begin{aligned}\mathbb{P}(R \notin \mathcal{R}_\varepsilon) &= \mathbb{P}(-\log p_r(R) \ge H(R)+\varepsilon) = \mathbb{P}(-\lambda\log p_r(R) \ge \lambda H(R) + \lambda\varepsilon) \\ &= \mathbb{P}\left(p_r^{-\lambda}(R) \ge \exp(\lambda H(R)+\lambda\varepsilon)\right) \\ &\le \frac{\mathbb{E}_R\left[p_r^{-\lambda}(R)\right]}{\exp(\lambda H(R)+\lambda\varepsilon)} = \frac{\sum_{r\in\mathcal{R}} p_r^{1-\lambda}(r)}{\exp(\lambda H(R)+\lambda\varepsilon)} = \frac{C_\lambda}{e^{\lambda\varepsilon}}. \end{aligned} \quad \text{(附录 A-48)}$$

解不等式 $-\log p_r(R) - H(R) \le \varepsilon$，可得 $\exp(-H(R)-\varepsilon) \le p_r(R)$。进一步有

$$1 \ge \mathbb{P}(R \in \mathcal{R}_\varepsilon) = \sum_{r\in\mathcal{R}_\varepsilon} p_r(r) \ge \sum_{r\in\mathcal{R}_\varepsilon} \exp(-H(R)-\varepsilon) = |\mathcal{R}_\varepsilon|\exp(-H(R)-\varepsilon).$$

在式 (附录 A-48) 中取 $C_\lambda/e^{\lambda\varepsilon} = \delta$，即 $\varepsilon = \frac{1}{\lambda}\log(C_\lambda/\delta)$，则有

$$\begin{aligned}|\mathcal{R}_\varepsilon| &\le \exp\left(H(R) + \frac{1}{\lambda}\log\left(\frac{C_\lambda}{\delta}\right)\right) = \exp\left(H(R) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \frac{1}{\lambda}\log\left(\frac{1}{e^{\lambda H(R)}}\sum_{r\in\mathcal{R}} p_r^{1-\lambda}(r)\right)\right) \\ &= \exp\left(H(R) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) - \frac{1}{\lambda}\lambda H(R) + \frac{1}{\lambda}\log\sum_{r\in\mathcal{R}} p_r^{1-\lambda}(r)\right) = \exp\left(H_{1-\lambda}(R) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right). \end{aligned}$$ ■

证明 (定理 3.3)**:** 假设 $R^W = r = \{l_i\}_{i=1}^{n+1} \in \mathcal{R}^W$，则有

$$\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)=l_U-\frac{1}{n}\sum_{i\neq U}l_i=l_U-\frac{1}{n}\sum_{i=1}^{n+1}l_i+\frac{1}{n}l_U=\frac{n+1}{n}l_U-\frac{n+1}{n}\bar{l}=\frac{n+1}{n}\left(l_U-\bar{l}\right),$$

其中 $\bar{l}=\frac{1}{n+1}\sum_{i=1}^{n+1}l_i$。容易验证 $\mathbb{E}_U\left[\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)\right]=0$。若 $l_U=b^{W,\widetilde{\mathbf{Z}}_l}=\sup_{i\in[1,n+1]}l_i$ 且对于任意 $i\neq U$ 有 $l_i=0$，则 $\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)$ 取最大值 $b^{W,\widetilde{\mathbf{Z}}_l}$。类似地，可证明 $\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)\geq -b^{W,\widetilde{\mathbf{Z}}_l}$，故 $\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)$ 满足 $b^{W,\widetilde{\mathbf{Z}}_l}$-次高斯性。假设 $R^W=r$ 且设 $\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)$ 对于 U 满足 σ_r-次高斯性，其中 $\sigma_r\in[0,b^{W,\widetilde{\mathbf{Z}}_l}]$，则对于任意 $t>0$，有

$$\mathbb{P}_U\left(\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)\geq t\right)\leq\exp\left(-\frac{t^2}{2(\sigma_r)^2}\right).$$

即对于任意 $\delta>0$ 与 $r\in\mathcal{R}^W$，若 $R^W=r$，则以置信度 $1-\delta$ 有 $\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)\leq\sigma_r\sqrt{2\log(1/\delta)}$。对 $r\in\mathcal{R}_\varepsilon^W$ 取联合上界，则当 $R^W\in\mathcal{R}_\varepsilon^W$ 时，有

$$\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)\leq\Sigma_{R^W}\sqrt{2\log(\left|\mathcal{R}_\varepsilon^W\right|/\delta)}.\tag{附录 A-49}$$

根据引理 附录 A.41，对于任意 $\delta>0$ 有

$$\mathbb{P}\left(R^W\notin\mathcal{R}_\varepsilon^W\right)\leq\delta,\qquad\left|\mathcal{R}_\varepsilon^W\right|\leq\exp\left(H_{1-\lambda}\left(R^W\right)+\frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right).\tag{附录 A-50}$$

取式 (附录 A-49) 与式 (附录 A-50) 的联合上界，则对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_l,U\right)\leq\Sigma_{R^W}\sqrt{2\log(2\left|\mathcal{R}_\varepsilon^W\right|/\delta)}\leq\Sigma_{R^W}\sqrt{2H_{1-\lambda}(R^W)+\frac{2}{\lambda}\log\left(\frac{1}{\delta}\right)+2\log\left(\frac{2}{\delta}\right)}.\ \blacksquare$$

证明 (定理 3.4)**:** 假设 $\widetilde{R}_\Delta^W=r=\{\Delta l_i\}_{i=1}^n\in\widetilde{\mathcal{R}}_\Delta^W$，则有

$$\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_s,\widetilde{U}\right)=L_{\overline{\mathbf{z}}}(W)-L_{\mathbf{z}}(W)=\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}}^W-L_{i,\widetilde{U}}^W=\frac{1}{n}\sum_{i=1}^{n}(-1)^{\widetilde{U}_i}\Delta l_i.$$

注意到 $\mathbb{E}_{\widetilde{U}_i}[(-1)^{\widetilde{U}_i}]=0$，通过对 $f(\widetilde{U})=\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_s,\widetilde{U}\right)$ 应用 McDiarmid 不等式，则对于任意 $t>0$，有

$$\mathbb{P}_{\widetilde{U}}\left(\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_s,\widetilde{U}\right)\geq t\right)\leq\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(2\Delta l_i/n)^2}\right).$$

即对于任意 $\delta>0$ 与 $r\in\widetilde{\mathcal{R}}_{\Delta,\varepsilon}^W$，若 $\widetilde{R}_\Delta^W=r$，则以置信度 $1-\delta$ 有

$$\overline{\mathrm{gen}}\left(W,\widetilde{\mathbf{Z}}_s,\widetilde{U}\right)\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\Delta l_i)^2}\sqrt{\frac{2\log(1/\delta)}{n}}.$$

对 $r \in \widetilde{\mathcal{R}}^{W}_{\Delta,\varepsilon}$ 取联合上界，则当 $\widetilde{R}^{W}_{\Delta} \in \widetilde{\mathcal{R}}^{W}_{\Delta,\varepsilon}$ 时，有

$$\overline{\mathrm{gen}}\left(W, \widetilde{\mathbf{Z}}_s, \widetilde{U}\right) \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta L_i^W\right)^2}\sqrt{\frac{2\log\left(\left|\widetilde{\mathcal{R}}^{W}_{\Delta,\varepsilon}\right|/\delta\right)}{n}}. \tag{附录 A-51}$$

根据引理 附录 A.41，对于任意 $\delta > 0$ 有

$$\mathbb{P}\left(\widetilde{R}^{W}_{\Delta} \notin \widetilde{\mathcal{R}}^{W}_{\Delta,\varepsilon}\right) \leq \delta, \qquad \left|\widetilde{\mathcal{R}}^{W}_{\Delta,\varepsilon}\right| \leq \exp\left(H_{1-\lambda}(\widetilde{R}^{W}_{\Delta}) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right). \tag{附录 A-52}$$

取式 (附录 A-51) 与式 (附录 A-52) 的联合上界，则对于任意 $\delta > 0$，以置信度 $1-\delta$ 有

$$\begin{aligned}\overline{\mathrm{gen}}\left(W, \widetilde{\mathbf{Z}}_s, \widetilde{U}\right) &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta L_i^W\right)^2}\sqrt{\frac{2\log\left(2\left|\widetilde{\mathcal{R}}^{W}_{\Delta,\varepsilon}\right|/\delta\right)}{n}} \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta L_i^W\right)^2}\sqrt{\frac{2H_{1-\lambda}(\widetilde{R}^{W}_{\Delta}) + \frac{2}{\lambda}\log\left(\frac{1}{\delta}\right) + 2\log\left(\frac{2}{\delta}\right)}{n}}. \end{aligned}$$ ■

引理 附录 A.42 给定任意 $\kappa > 0$，$\lambda \in (0,1)$ 与 $\delta > 0$，则以置信度 $1-\delta$，有

$$\frac{1}{n}\sum_{i=1}^{n} L^{W,\kappa}_{i,1-\widetilde{U}_i} - (1+C_i)L^{W,\kappa}_{i,\widetilde{U}_i} \leq \frac{H_{1-\lambda}(\widetilde{R}^{W,\kappa}) + \frac{1}{\lambda}\log(1/\delta) + \log(4/\delta)}{n\eta},$$

其中 $\widetilde{R}^{W,\kappa} = \{L^{W,\kappa}_{i,0}, L^{W,\kappa}_{i,1}\}_{i=1}^{n}$ 且对于任意 $i \in [1,n]$，定义

$$\eta \in \left(0, \frac{\log 2}{2\kappa}\right), \quad C_i = -\frac{\log\left(2 - e^{2\eta\widehat{L}^{W,\kappa}_i}\right)}{2\eta\widehat{L}^{W,\kappa}_i} - 1, \quad \widehat{L}^{W,\kappa}_i = \max\left(L^{W,\kappa}_{i,0}, L^{W,\kappa}_{i,1}\right).$$

证明: 假设 $\widetilde{R}^{W,\kappa} = r = \{l_{i,0}, l_{i,1}\}_{i=1}^{n} \in \widetilde{\mathcal{R}}^{W,\kappa}$，则 $C_i, i \in [1,n]$ 可视为由 r 确定的常数，故有

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} l_{i,1-\widetilde{U}_i} - (1+C_i)l_{i,\widetilde{U}_i} \geq t\right) \\ &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(1+\frac{C_i}{2}\right)\left(l_{i,1-\widetilde{U}_i} - l_{i,\widetilde{U}_i}\right) - \frac{C_i}{2}l_{i,1-\widetilde{U}_i} - \frac{C_i}{2}l_{i,\widetilde{U}_i} \geq t\right) \\ &= \mathbb{P}\left(\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)l_{i,1} - C_i l_{i,1}\right) + \frac{1}{2n}\sum_{i=1}^{n}\left(-(-1)^{\widetilde{U}_i}(2+C_i)l_{i,0} - C_i l_{i,0}\right) \geq t\right) \\ &\leq \mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i) - C_i\right)l_{i,I} + \frac{1}{2n}\sum_{i=1}^{n}\left(-(-1)^{\widetilde{U}_i}(2+C_i) - C_i\right)l_{i,1-I}\right\} \geq t\right)\end{aligned}$$

$$\leq \mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right\}+\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(-(-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,1-I}\right\}\geq t\right)$$

$$\leq \inf_{\gamma\in(0,1)}\mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right\}\geq \gamma t\right)$$

$$+\mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(-(-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,1-I}\right\}\geq (1-\gamma)t\right).$$

注意到上式中的两个概率事件拥有相同的边际分布，故有

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}l_{i,1-\widetilde{U}_i}-(1+C_i)l_{i,\widetilde{U}_i}\geq t\right)\leq \inf_{\gamma\in(0,1)}\mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right\}\geq \gamma t\right)$$

$$+\mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right\}\geq (1-\gamma)t\right). \tag{附录 A-53}$$

由于 C_i 可视为常数，故其与 $\widetilde{U}$ 独立。对于任意 $I\in\{0,1\}$，$t>0$ 与 $\eta>0$，应用 Markov 不等式可得

$$\begin{aligned}\mathbb{P}\left(\frac{1}{2n}\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\geq t\right)&=\mathbb{P}\left(\exp\left(\eta\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right)\geq e^{2\eta nt}\right)\\&\leq e^{-2\eta nt}\mathbb{E}_{\widetilde{U}}\left[\exp\left(\eta\sum_{i=1}^{n}\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right)\right]\\&=e^{-2\eta nt}\prod_{i=1}^{n}\mathbb{E}_{\widetilde{U}_i}\left[\exp\left(\eta\left((-1)^{\widetilde{U}_i}(2+C_i)-C_i\right)l_{i,I}\right)\right]\\&=e^{-2\eta nt}\prod_{i=1}^{n}\frac{e^{-2\eta l_{i,I}(1+C_i)}+e^{2\eta l_{i,I}}}{2}.\end{aligned}$$

通过构造合适的 η 与 C_i 值，可使 $e^{-2\eta l_{i,I}(1+C_i)}+e^{2\eta l_{i,I}}\leq 2$ 对于任意 $i\in[1,n]$ 与 $I\in\{0,1\}$ 均成立。注意到 $e^{2\eta l_{i,I}}\leq 2$ 隐含了 $2\eta l_{i,I}\leq \log 2<1$。此外，$e^{-2\eta l_{i,I}(1+C_i)}+e^{2\eta l_{i,I}}$ 随 C_i 增大而单调递减。求解此不等式可得 $C_i\geq -\log(2-e^{2\eta l_{i,I}})/2\eta l_{i,I}-1$。容易验证此下界随 $l_{i,I}$ 增大而单调递增。因此，若选择

$$C_i\geq -\frac{\log(2-e^{2\eta\max(l_{i,0},l_{i,1})})}{2\eta\max(l_{i,0},l_{i,1})}-1,$$

则可保证对于任意 $i\in[1,n]$，均有 $e^{-2\eta l_{i,I}(1+C_i)}+e^{2\eta l_{i,I}}\leq 2$。代入上式则有

$$\mathbb{P}\left(\frac{1}{2n}\sum_{i=1}^{n}\Big((-1)^{\widetilde{U}_i}(2+C_i)-C_i\Big)l_{i,I}\geq t\right)\leq e^{-2\eta nt}. \tag{附录 A-54}$$

将上式对 $I\in\{0,1\}$ 取联合上界，可得

$$\mathbb{P}\left(\sup_{I\in\{0,1\}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\Big((-1)^{\widetilde{U}_i}(2+C_i)-C_i\Big)l_{i,I}\right\}\geq t\right)\leq 2e^{-2\eta nt}. \tag{附录 A-55}$$

将上式代入式 (附录 A-53)，可得

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}l_{i,1-\widetilde{U}_i}-(1+C_i)l_{i,\widetilde{U}_i}\geq t\right)\leq\inf_{\gamma\in(0,1)}2e^{-2\eta\gamma nt}+2e^{-2\eta(1-\gamma)nt}=4e^{-\eta nt}. \tag{附录 A-56}$$

令 δ 等于上式右侧，则对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$\frac{1}{n}\sum_{i=1}^{n}l_{i,1-\widetilde{U}_i}-(1+C_i)l_{i,\widetilde{U}_i}\leq\frac{\log(4/\delta)}{n\eta}.$$

根据引理 附录 A.41，对于任意 $\delta>0$ 有

$$\mathbb{P}\Big(\widetilde{R}^{W,\kappa}\notin\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}\Big)\leq\delta,\qquad\Big|\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}\Big|\leq\exp\left(H_{1-\lambda}(\widetilde{R}^{W,\kappa})+\frac{1}{\lambda}\log\left(\frac{1}{\delta}\right)\right). \tag{附录 A-57}$$

注意到由于 $\widetilde{U}$ 的存在，$\widetilde{R}^{W,\kappa}$ 的边际分布具有对称性，即若 $\{l_{i,0},l_{i,1}\}_{i=1}^{n}\in\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}$，则同样有 $\{l_{i,1},l_{i,0}\}_{i=1}^{n}\in\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}$。由于式 (附录 A-55) 中的上界对于 $I\in\{0,1\}$ 同时成立，$\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}$ 的等效大小可除以 2。对 $r\in\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}$ 取联合上界，则对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,\kappa}-(1+C_i)L_{i,\widetilde{U}_i}^{W,\kappa}\leq\frac{\log\Big(2\Big|\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}\Big|/\delta\Big)}{n\eta}. \tag{附录 A-58}$$

取式 (附录 A-57) 与式 (附录 A-58) 的联合上界，则对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,\kappa}-(1+C_i)L_{i,\widetilde{U}_i}^{W,\kappa}&\leq\frac{\log\Big(\Big|\widetilde{\mathcal{R}}_{\varepsilon}^{W,\kappa}\Big|\Big)+\log(4/\delta)}{n\eta}\\&\leq\frac{H_{1-\lambda}(\widetilde{R}^{W,\kappa})+\frac{1}{\lambda}\log(1/\delta)+\log(4/\delta)}{n\eta}.\end{aligned}$$

■

证明 (定理 3.5)**:** 注意到当 $\kappa\geq B^{W,\widetilde{\mathbf{z}}_s}$，对于任意 $i\in[1,n]$ 与 $I\in\{0,1\}$ 有 $L_{i,I}^{W,\kappa}=L_{i,I}^{W}$，故该定理可直接通过引理 附录 A.42 得到。 ■

证明 (定理 3.6)**:** 根据验证误差的定义，有以下误差分解：

$$\overline{\mathrm{gen}}\Big(W,\widetilde{\mathbf{Z}}_s,\widetilde{U}\Big)=\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,\kappa}-(1+C_i)L_{i,\widetilde{U}_i}^{W,\kappa}+\frac{1}{n}\sum_{i=1}^{n}C_iL_{i,\widetilde{U}_i}^{W,\kappa}+\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,-\kappa}-L_{i,\widetilde{U}_i}^{W,-\kappa}. \tag{附录 A-59}$$

应用引理 附录 A.42，则对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,\kappa}-(1+C_i)L_{i,\widetilde{U}_i}^{W,\kappa}\le\frac{H_{1-\lambda_1}(\widetilde{R}^{W,\kappa})+\frac{1}{\lambda_1}\log(1/\delta)+\log(4/\delta)}{n\eta},$$

取 $\eta=\frac{\gamma\log 2}{2\kappa}<\frac{\log 2}{2\kappa}$，则可得

$$\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,\kappa}-(1+C_i)L_{i,\widetilde{U}_i}^{W,\kappa}\le\frac{2\kappa\Big(H_{1-\lambda_1}(\widetilde{R}^{W,\kappa})+\frac{1}{\lambda_1}\log(1/\delta)+\log(4/\delta)\Big)}{n\gamma\log 2}. \tag{附录 A-60}$$

通过将定理 3.4 的证明过程应用于 $\widetilde{R}^{W,-\kappa}$，可得对于任意 $\delta>0$，以置信度 $1-\delta$ 有

$$\frac{1}{n}\sum_{i=1}^{n}L_{i,1-\widetilde{U}_i}^{W,-\kappa}-L_{i,\widetilde{U}_i}^{W,-\kappa}\le\sqrt{\frac{2}{n}\sum_{i=1}^{n}\Big(\Delta L_i^{W,-\kappa}\Big)^2}\sqrt{\frac{H_{1-\lambda_2}(\widetilde{R}_{\Delta,d}^{W,-\kappa})+\frac{1}{\lambda_2}\log(1/\delta)+\log(2/\delta)}{n}}. \tag{附录 A-61}$$

取式 (附录 A-60) 与式 (附录 A-61) 的联合上界并代入式 (附录 A-59)，可得

$$\begin{aligned}\overline{\mathrm{gen}}\Big(W,\widetilde{\mathbf{Z}}_s,\widetilde{U}\Big)\le{}&\frac{1}{n}\sum_{i=1}^{n}C_iL_{i,\widetilde{U}_i}^{W,\kappa}+\frac{2\kappa\Big(H_{1-\lambda_1}(\widetilde{R}^{W,\kappa})+\frac{1}{\lambda_1}\log(2/\delta)+\log(8/\delta)\Big)}{n\gamma\log 2}\\&+\sqrt{\frac{2}{n}\sum_{i=1}^{n}\Big(\Delta L_i^{W,-\kappa}\Big)^2}\sqrt{\frac{H_{1-\lambda_2}(\widetilde{R}_{\Delta}^{W,-\kappa})+\frac{1}{\lambda_2}\log(2/\delta)+\log(4/\delta)}{n}}.\end{aligned}$$ ■

A.4 第 4 章定理补充证明

引理 附录 A.43 设 $X_1,\cdots,X_n$ 为独立随机变量，则对于任意随机变量 Y，有

$$I(X_1;Y)+\cdots+I(X_n;Y)\le I(X_1,\cdots,X_n;Y).$$

证明: 因为 $X_1,\cdots,X_n$ 互相独立，故 $I(X_2,\cdots,X_n;X_1)=0$ 且

$$\begin{aligned}I(X_1,\cdots,X_n;Y)&=I(X_1;Y)+I(X_2,\cdots,X_n;Y|X_1)\\&=I(X_1;Y)+I(X_2,\cdots,X_n;Y)+I(X_2,\cdots,X_n;X_1|Y)-I(X_2,\cdots,X_n;X_1)\\&=I(X_1;Y)+I(X_2,\cdots,X_n;Y)+I(X_2,\cdots,X_n;X_1|Y)\\&\ge I(X_1;Y)+I(X_2,\cdots,X_n;Y).\end{aligned}$$

通过重复上述递推过程，引理得证。类似结论同样适用于解构互信息或条件互信息。 ■

引理 附录 A.44 设 $X\sim N(0,\Sigma)$，Y 为任意满足 $\mathrm{Cov}_Y[Y]=\Sigma$ 且期望为 0 的随机向量，则

$H(Y) \leq H(X)$.

证明: 根据条件 $\mathrm{Cov}_Y[Y] = \Sigma$，设 X 与 Y 均为 d 维变量，则

$$\int p_Y(x) x^\top \Sigma^{-1} x \,\mathrm{d}x = \int p_Y(x) \mathrm{tr}(xx^\top \Sigma^{-1}) \,\mathrm{d}x = \mathrm{tr}(\Sigma\Sigma^{-1}) = d = \int p_X(x) x^\top \Sigma^{-1} x \,\mathrm{d}x.$$

因此，有

$$\begin{aligned}
0 \leq D(P_Y \,\|\, P_X) &= \int p_Y(x) \log \frac{p_Y(x)}{p_X(x)} \,\mathrm{d}x \\
&= -H(Y) - \int p_Y(x) \log p_X(x) \,\mathrm{d}x \\
&= -H(Y) + \int p_Y(x) \left(\frac{d}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma| + \frac{1}{2} x^\top \Sigma^{-1} x \right) \mathrm{d}x \\
&= -H(Y) + \int p_X(x) \left(\frac{d}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma| + \frac{1}{2} x^\top \Sigma^{-1} x \right) \mathrm{d}x \\
&= -H(Y) - H(X).
\end{aligned}$$

■

证明 (定理 4.2)**:** 根据期望泛化误差的定义，给定独立同分布样本数据 $Z'_{1:m} \sim \mu^m$，则有

$$\begin{aligned}
|\overline{\mathrm{gen}}| &= \left| \mathbb{E}_{W,Z}[L(W) - L_Z(W)] \right| \\
&= \left| \mathbb{E}_{W,Z'_{1:m}}[\ell(W, Z'_{1:m})] - \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W,Z_u}[\ell(W, Z_u)] \right| \\
&\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left| \mathbb{E}_{W,Z'_{1:m}}[\ell(W, Z'_{1:m})] - \mathbb{E}_{W,Z_u}[\ell(W, Z_u)] \right|. \qquad \text{(附录 A-62)}
\end{aligned}$$

由于 $\ell(\cdot,\cdot) \in [0,1]$，可得 $\ell(W, Z'_{1:m})$ 满足 $\frac{1}{2}$-次高斯条件。对于任意 $u \in \mathsf{P}_n^m$，由于 Z_u 同样由独立同分布样本构成，$Z'_{1:m}$ 可视作 Z_u 的一个独立拷贝。通过在引理 附录 A.7 中取 $f(W, Z_u) = \ell(W, Z_u)$，可得

$$\left| \mathbb{E}_{W,Z'_{1:m}}[\ell(W, Z'_{1:m})] - \mathbb{E}_{W,Z_u}[\ell(W, Z_u)] \right| \leq \sqrt{\frac{1}{2} I(W; Z_u)}.$$

将上述不等式代入式 (附录 A-62)，则有

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left| \mathbb{E}_{W,Z'_{1:m}}[\ell(W, Z'_{1:m})] - \mathbb{E}_{W,Z_u}[\ell(W, Z_u)] \right| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{\frac{1}{2} I(W; Z_u)}.$$

由于 $I(W; Z_u)$ 取值与 Z_u 中的样本顺序无关，可得

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{\frac{1}{2} I(W; Z_u)} = \frac{1}{|\mathsf{C}_n^m|} \sum_{u \in \mathsf{C}_n^m} \sqrt{\frac{1}{2} I(W; Z_u)}.$$

则对于任意 $k \in \left[1, \frac{n}{m}\right]$，有

$$
\begin{aligned}
|\overline{\mathrm{gen}}| &\le \frac{1}{|\mathsf{C}_n^m|}\sum_{u\in\mathsf{C}_n^m}\sqrt{\frac{1}{2}I(W;Z_u)} = \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\sqrt{\frac{1}{2}I(W;(Z_u)_v)} \\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{k}\underbrace{\left(\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\sqrt{\frac{1}{2}I(W;(Z_u)_v)} + \cdots + \frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\sqrt{\frac{1}{2}I(W;(Z_u)_v)}\right)}_{\times k} \\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\frac{1}{k}\left(\sqrt{\frac{1}{2}I(W;((Z_u)_v)_{1:m})} + \cdots + \sqrt{\frac{1}{2}I(W;((Z_u)_v)_{(k-1)m+1:km})}\right) \\
&\le \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\sqrt{\frac{1}{2k}\Big(I(W;((Z_u)_v)_{1:m}) + \cdots + I(W;((Z_u)_v)_{(k-1)m+1:km})\Big)} \qquad \text{(附录 A-63)} \\
&\le \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\sqrt{\frac{1}{2k}I(W;(Z_u)_v)} = \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\sqrt{\frac{1}{2k}I(W;Z_u)} \qquad \text{(附录 A-64)}
\end{aligned}
$$

其中式 (附录 A-63) 可通过对平方根函数应用 Jensen 不等式得到，式 (附录 A-64) 可通过引理 附录 A.43 得到。定理得证。 ■

证明 (定理 4.3)**:** 由于函数 $d_\gamma(\cdot\,\|\,\cdot)$ 的联合凸性，应用 Jensen 不等式可得

$$
\begin{aligned}
d(L_n\,\|\,L) = \sup_\gamma d_\gamma(L_n\,\|\,L) &= \sup_\gamma d_\gamma\left(\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,Z_u}[\ell(W,Z_u)]\,\middle\|\,\mathbb{E}_W[L(W)]\right) \\
&\le \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,Z_u}\Big[d_\gamma(\ell(W,Z_u)\,\|\,L(W))\Big]. \qquad \text{(附录 A-65)}
\end{aligned}
$$

给定独立同分布样本 $Z'_{1:m}\sim\mu^m$。对于任意 $u\in\mathsf{P}_n^m$，通过在引理 附录 A.9 中取 $P = P_{W,Z_u}$，$Q = P_W P_{Z'_{1:m}}$ 与 $f = d_\gamma$，可得

$$
I(W;Z_u) \ge \mathbb{E}_{W,Z_u}[d_\gamma(\ell(W,Z_u)\,\|\,L(W))] - \log\mathbb{E}_{W,Z'_{1:m}}\left[e^{d_\gamma\left(\ell(W,Z'_{1:m})\,\|\,L(W)\right)}\right]. \qquad \text{(附录 A-66)}
$$

对于任意 $w\in\mathcal{W}$，可得 $\mathbb{E}_{Z'_{1:m}}[\ell(w,Z'_{1:m})] = L(w)$。注意到 $\ell(\cdot,\cdot)\in[0,1]$，则通过引理 附录 A.12，对于任意 $\gamma\in\mathbb{R}$ 有 $\mathbb{E}_{Z'_{1:m}}\left[e^{d_\gamma\left(\ell(w,Z'_{1:m})\,\|\,L(w)\right)}\right]\le 1$。由于 W 独立于 $Z'_{1:m}$，故

$$
\mathbb{E}_{W,Z'_{1:m}}\left[e^{d_\gamma\left(\ell(W,Z'_{1:m})\,\|\,L(W)\right)}\right] = \mathbb{E}_W\left[\mathbb{E}_{Z'_{1:m}}\left[e^{d_\gamma\left(\ell(W,Z'_{1:m})\,\|\,L(W)\right)}\right]\right] \le 1.
$$

代入式 (附录 A-66)，可得 $\mathbb{E}_{W,Z_u}[d_\gamma(\ell(W,Z_u)\,\|\,L(W))]\le I(W;Z_u)$。代入式 (附录 A-65)，有

$$d(L_n \,\|\, L) \le \sup_{\gamma} \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{W,Z_u}\Big[d_\gamma(\ell(W,Z_u) \,\|\, L(W))\Big] \le \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} I(W;Z_u).$$

类似地，由于 $I(W;Z_u)$ 与 Z_u 中样本顺序无关，对于任意 $k \in \left[1, \frac{n}{m}\right]$ 有

$$\begin{aligned}
d(L_n \,\|\, L) &\le \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} I(W;Z_u) = \frac{1}{|\mathsf{C}_n^m|} \sum_{u\in\mathsf{C}_n^m} I(W;Z_u) = \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} \frac{1}{|\mathsf{P}_{km}^m|} \sum_{v\in\mathsf{P}_{km}^m} I(W;(Z_u)_v) \\
&= \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} \frac{1}{k} \underbrace{\left(\frac{1}{|\mathsf{P}_{km}^m|} \sum_{v\in\mathsf{P}_{km}^m} I(W;(Z_u)_v) + \cdots + \frac{1}{|\mathsf{P}_{km}^m|} \sum_{v\in\mathsf{P}_{km}^m} I(W;(Z_u)_v) \right)}_{\times k} \\
&= \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} \frac{1}{|\mathsf{P}_{km}^{km}|} \sum_{v\in\mathsf{P}_{km}^{km}} \Big(I(W;((Z_u)_v)_{1:m}) + \cdots + I(W;((Z_u)_v)_{(k-1)m+1:km}) \Big) \\
&\le \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} \frac{1}{|\mathsf{P}_{km}^{km}|} \sum_{v\in\mathsf{P}_{km}^{km}} I(W;(Z_u)_v) = \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} I(W;Z_u) \qquad \text{(附录 A-67)}
\end{aligned}$$

其中式 (附录 A-67) 可通过引理 附录 A.43 得到。进一步考虑 $L_n = 0$ 的情形，有

$$d(L_n \,\|\, L) = d(0 \,\|\, L) = \log\left(\frac{1}{1-L}\right) \ge L. \qquad \blacksquare$$

证明 (定理 4.4)**:** 对于任意 $u \in \mathsf{P}_n^{km+m}$，应用互信息的链式分解，可得

$$\begin{aligned}
I(W;Z_u) &= \sum_{i=1}^{k+1} I(W;(Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}) \\
&= \sum_{i=1}^{k+1} I(W,(Z_u)_{im+1:(k+1)m};(Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}) \\
&\quad - I((Z_u)_{(i-1)m+1:im};(Z_u)_{im+1:(k+1)m} | (Z_u)_{1:(i-1)m}, W) \\
&\le \sum_{i=1}^{k+1} I(W,(Z_u)_{im+1:(k+1)m};(Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}) \\
&= \sum_{i=1}^{k+1} I(W;(Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m},(Z_u)_{im+1:(k+1)m}) \\
&\quad + I((Z_u)_{(i-1)m+1:im};(Z_u)_{im+1:(k+1)m} | (Z_u)_{1:(i-1)m}) \\
&= \sum_{i=1}^{k+1} I(W;(Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m},(Z_u)_{im+1:(k+1)m}).
\end{aligned}$$

类似地，对于任意 $u \in \mathsf{C}_n^{km+m}$，在上述不等式中取 $Z_u = (Z_u)_v$，有

$$I(W;Z_u) = \frac{1}{\left|\mathsf{P}_{km+m}^m\right|} \sum_{v\in\mathsf{P}_{km+m}^m} I(W;Z_u \setminus (Z_u)_v) + I(W;(Z_u)_v | Z_u \setminus (Z_u)_v)$$

$$
\begin{aligned}
&= \frac{1}{\left|\mathsf{P}_{km+m}^{km}\right|} \sum_{v\in\mathsf{P}_{km+m}^{km}} I(W;(Z_u)_v) + \frac{1}{k+1}\sum_{i=1}^{k+1} I(W;(Z_u)_v | Z_u \setminus (Z_u)_v) \\
&= \frac{1}{\left|\mathsf{P}_{km+m}^{km}\right|} \sum_{v\in\mathsf{P}_{km+m}^{km}} I(W;(Z_u)_v) \\
&\quad + \frac{1}{\left|\mathsf{P}_{km+m}^{km+m}\right|} \sum_{v\in\mathsf{P}_{km+m}^{km+m}} \frac{1}{k+1}\sum_{i=1}^{k+1} I(W;((Z_u)_v)_{(i-1)m+1:im} | Z_u \setminus ((Z_u)_v)_{(i-1)m+1:im}) \\
&\geq \frac{1}{\left|\mathsf{P}_{km+m}^{km}\right|} \sum_{v\in\mathsf{P}_{km+m}^{km}} I(W;(Z_u)_v) + \frac{1}{\left|\mathsf{P}_{km+m}^{km+m}\right|} \sum_{v\in\mathsf{P}_{km+m}^{km+m}} \frac{1}{k+1} I(W;(Z_u)_v).
\end{aligned}
$$

将上式变形可得

$$
\frac{1}{k+1} I(W;Z_u) \geq \frac{1}{k\left|\mathsf{C}_{km+m}^{km}\right|} \sum_{v\in\mathsf{C}_{km+m}^{km}} I(W;(Z_u)_v).
$$

因此，对凹函数 φ 应用 Jensen 不等式，可得

$$
\begin{aligned}
\frac{1}{\left|\mathsf{C}_n^{km+m}\right|} \sum_{u\in\mathsf{C}_n^{km+m}} \varphi\left(\frac{1}{2(k+1)} I(W;Z_u)\right) &\geq \frac{1}{\left|\mathsf{C}_n^{km+m}\right|} \sum_{u\in\mathsf{C}_n^{km+m}} \varphi\left(\frac{1}{2k\left|\mathsf{C}_{km+m}^{km}\right|} \sum_{v\in\mathsf{C}_{km+m}^{km}} I(W;(Z_u)_v)\right) \\
&\geq \frac{1}{\left|\mathsf{C}_n^{km+m}\right|} \sum_{u\in\mathsf{C}_n^{km+m}} \frac{1}{\left|\mathsf{C}_{km+m}^{km}\right|} \sum_{v\in\mathsf{C}_{km+m}^{km}} \varphi\left(\frac{1}{2k} I(W;(Z_u)_v)\right) \\
&= \frac{1}{\left|\mathsf{C}_n^{km}\right|} \sum_{u\in\mathsf{C}_n^{km}} \varphi\left(\frac{1}{2k} I(W;Z_u)\right).
\end{aligned}
$$

■

证明 (定理 4.6)**:** 根据期望泛化误差的定义，可得

$$
\begin{aligned}
|\overline{\mathrm{gen}}| = \left|\mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L_{\widetilde{\mathbf{Z}}_{\bar{S}}}(W) - L_{\widetilde{\mathbf{Z}}_S}(W)\right]\right| &\leq \mathbb{E}_{\widetilde{\mathbf{Z}}}\left|\mathbb{E}_{W,S|\widetilde{\mathbf{Z}}}\left[L_{\widetilde{\mathbf{Z}}_{\bar{S}}}(W) - L_{\widetilde{\mathbf{Z}}_S}(W)\right]\right| \\
&\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{\widetilde{\mathbf{z}}}\left|\mathbb{E}_{W,S_u|\widetilde{\mathbf{z}}}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right]\right|. \qquad \text{(附录 A-68)}
\end{aligned}
$$

设 S' 为 S 的一个独立拷贝，使得 $S' \perp\!\!\!\perp W|\widetilde{\mathbf{Z}} = \tilde{z}$。对于任意 $u \in \mathsf{P}_n^m$，在引理 附录 A.9 中取 $P = P_{W,S_u|\tilde{z}}$，$Q = P_{W|\tilde{z}} P_{S_u}$ 与 $f(W,S_u) = L_u^{\bar{S}_u} - L_u^{S_u}$，可得

$$
\widetilde{F}(W;S_u) \geq \sup_{t\in\mathbb{R}}\left\{\mathbb{E}_{W,S_u|\tilde{z}}\left[t\left(L_u^{\bar{S}_u} - L_u^{S_u}\right)\right] - \log \mathbb{E}_{W,S_u'|\tilde{z}}\left[\exp\left(t\left(L_u^{\bar{S}_u'} - L_u^{S_u'}\right)\right)\right]\right\}. \quad \text{(附录 A-69)}
$$

注意到 $f(W,S_u') \in [-1,1]$ 且 $\mathbb{E}_{W,S_u'|\tilde{z}}[f(W,S_u')] = 0$，则其满足 1-次高斯性。根据定义，可得 $\mathbb{E}_{W,S_u'|\tilde{z}}\left[\exp\left(t\left(L_u^{\bar{S}_u'} - L_u^{S_u'}\right)\right)\right] \leq e^{\frac{t^2}{2}}$。代入式 (附录 A-69)，可得 $\left|\mathbb{E}_{W,S_u|\tilde{z}}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right]\right| \leq \sqrt{2\widetilde{F}(W;S_u)}$。将其代入式 (附录 A-68)，可得

$$|\overline{\mathrm{gen}}| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{Z}}}\left|\mathbb{E}_{W,S_u|\widetilde{\mathbf{Z}}}\left[L_u^{\bar{S}_u}-L_u^{S_u}\right]\right| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{Z}}}\sqrt{2I^{\widetilde{\mathbf{Z}}}(W;S_u)}.$$

通过与定理 4.2 相似的证明步骤，可得对于任意 $k\in\left[1,\frac{n}{m}\right]$，有

$$|\overline{\mathrm{gen}}| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{Z}}}\sqrt{2I^{\widetilde{\mathbf{Z}}}(W;S_u)} \le \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{\widetilde{\mathbf{Z}}}\sqrt{\frac{2}{k}I^{\widetilde{\mathbf{Z}}}(W;S_u)}. \qquad \blacksquare$$

证明 (定理 4.7)**:** 通过 Jensen 不等式以及 $d_\gamma(\cdot\|\cdot)$ 的联合凸性，可得

$$\begin{aligned}
d\left(L_n\middle\|\frac{L_n+L}{2}\right) &= \sup_\gamma d_\gamma\left(L_n\middle\|\frac{L_n+L}{2}\right)\\
&\le \sup_\gamma \mathbb{E}_{\widetilde{\mathbf{Z}}}\left[d_\gamma\left(\mathbb{E}_{W,S|\widetilde{\mathbf{Z}}}\left[L_{\widetilde{\mathbf{Z}}_S}(W)\right]\middle\|\mathbb{E}_{W,S|\widetilde{\mathbf{Z}}}\left[\frac{L_{\widetilde{\mathbf{Z}}_S}(W)+L_{\widetilde{\mathbf{Z}}_{\bar{S}}}(W)}{2}\right]\right)\right]\\
&= \sup_\gamma \mathbb{E}_{\widetilde{\mathbf{Z}},\Phi}\left[d_\gamma\left(\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,S_u|\widetilde{\mathbf{Z}},\Phi_u}\left[L_u^{S_u}\right]\middle\|\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W|\widetilde{\mathbf{Z}},\Phi_u}\left[\frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right]\right)\right]\\
&\le \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{Z}},\Phi_u}\left[d_\gamma\left(\mathbb{E}_{W,S_u|\widetilde{\mathbf{Z}},\Phi_u}\left[L_u^{S_u}\right]\middle\|\mathbb{E}_{W|\widetilde{\mathbf{Z}},\Phi_u}\left[\frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right]\right)\right]\\
&\le \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{Z}},\Phi_u}\mathbb{E}_{W,S_u|\widetilde{\mathbf{Z}},\Phi_u}\left[d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right)\right]. \qquad \text{(附录 A-70)}
\end{aligned}$$

设 S' 为 S 的一个独立拷贝，满足 $S'\perp\!\!\!\perp W|\widetilde{\mathbf{Z}}=\tilde{z}$。对于任意 $u\in\mathsf{P}_n^m$，在引理 附录 A.9 中取 $P=P_{W,S_u|\tilde{z},\varphi_u}$，$Q=P_{W|\tilde{z},\varphi_u}P_{S_u|\varphi_u}$ 与 $f(W,S_u)=d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\varphi_u^+}+L_u^{\varphi_u^-}}{2}\right)$，可得

$$\widetilde{F}^{,\varphi_u}(W;S_u) \ge \mathbb{E}_{W,S_u|\tilde{z},\varphi_u}\left[d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\varphi_u^+}+L_u^{\varphi_u^-}}{2}\right)\right]-\log\mathbb{E}_{W,S'_u|\tilde{z},\varphi_u}\left[\exp\left(d_\gamma\left(L_u^{S'_u}\middle\|\frac{L_u^{\varphi_u^+}+L_u^{\varphi_u^-}}{2}\right)\right)\right]. \qquad \text{(附录 A-71)}$$

注意到 $\mathbb{E}_{S'_u|\varphi_u}\left[L_u^{S'_u}\right]=\frac{L_u^{\varphi_u^+}+L_u^{\varphi_u^-}}{2}\in[0,1]$。进一步地，应用引理 附录 A.12，可得对于任意 $\gamma\in\mathbb{R}$，有 $\mathbb{E}_{W,S'_u|\tilde{z},\varphi_u}\left[\exp\left(d_\gamma\left(L_u^{S'_u}\middle\|\frac{L_u^{\varphi_u^+}+L_u^{\varphi_u^-}}{2}\right)\right)\right]\le 1$。将其代入式 (附录 A-71)，可得

$$\mathbb{E}_{W,S_u|\tilde{z},\varphi_u}\left[d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\varphi_u^+}+L_u^{\varphi_u^-}}{2}\right)\right]\le\widetilde{F}^{,\varphi_u}(W;S_u).$$

将上式代入式 (附录 A-70)，可得

$$
\begin{aligned}
d\left(L_n \middle\| \frac{L_n+L}{2}\right) &\le \sup_{\gamma} \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{\widetilde{\mathbf{Z}},\Phi_u} \mathbb{E}_{W,S_u|\widetilde{\mathbf{Z}},\Phi_u}\left[d_\gamma\left(L_u^{S_u} \middle\| \frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right)\right] \\
&\le \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{\widetilde{\mathbf{Z}},\Phi_u}\left[I^{\widetilde{\mathbf{Z}},\Phi_u}(W;S_u)\right] \le \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \left(I(W;S_u|\widetilde{\mathbf{Z}},\Phi_u)+I(W;\Phi_u|\widetilde{\mathbf{Z}})\right) \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} I(W;S_u,\Phi_u|\widetilde{\mathbf{Z}}) = \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} I(W;S_u|\widetilde{\mathbf{Z}}).
\end{aligned}
$$

通过与定理 4.3 相似的证明步骤，可得对于任意 $k \in \left[1, \frac{n}{m}\right]$ 有

$$
d\left(L_n \middle\| \frac{L_n+L}{2}\right) \le \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} I(W;S_u|\widetilde{\mathbf{Z}}) \le \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} I(W;S_u|\widetilde{\mathbf{Z}}).
$$

当 $L_n = 0$ 时，可得 $d\left(L_n \middle\| \frac{L_n+L}{2}\right) = d\left(0 \middle\| \frac{L}{2}\right) \ge \frac{L}{2}$。 ■

证明 (定理 4.8)**:** 该定理的证明方法与定理 4.4 相同，仅需将证明过程中的互信息度量 $I(W;Z_u)$ 替换为解构互信息 $\widetilde{I}(W;S_u)$ 即得证。 ■

证明 (定理 4.10)**:** 根据加权泛化误差的定义，有

$$
\begin{aligned}
L-(1+C_1)L_n &= \mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L_{\widetilde{\mathbf{Z}}_{\bar{S}}}(W)-(1+C_1)L_{\widetilde{\mathbf{Z}}_S}(W)\right] \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{W,S_u,\widetilde{\mathbf{Z}},\Phi_u}\left[L_u^{\bar{S}_u}-(1+C_1)L_u^{S_u}\right] \qquad \text{(附录 A-72)} \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{\widetilde{\mathbf{Z}},\Phi_u} \mathbb{E}_{W,S_u|\widetilde{\mathbf{Z}},\Phi_u}\left[L_u^{\bar{S}_{u_1}\otimes\Phi_u}-(1+C_1)L_u^{S_{u_1}\otimes\Phi_u}\right]
\end{aligned}
$$

设 S' 为 S 的独立拷贝。对于任意 $u \in \mathsf{P}_n^m$，通过在引理 附录 A.9 中取 $P = P_{W,S_{u_1}|\widetilde{z},\varphi_u}$，$Q = P_{W|\widetilde{z},\varphi_u} P_{S_{u_1}}$ 与 $f(W,S_{u_1}) = L_u^{\bar{S}_{u_1}\otimes\varphi_u} - C_1 L_u^{S_{u_1}\otimes\varphi_u}$，可得

$$
\begin{aligned}
\widetilde{I}^{\widetilde{z},\varphi_u}(W;S_{u_1}) &\ge \sup_{C_2>0}\left\{\mathbb{E}_{W,S_{u_1}|\widetilde{z},\varphi_u}\left[C_2\left(L_u^{\bar{S}_u}-(1+C_1)L_u^{S_u}\right)\right] - \log \mathbb{E}_{W,S'_{u_1}|\widetilde{z},\varphi_u}\left[e^{C_2\left(L_u^{\bar{S}'_{u_1}\otimes\varphi_u}-(1+C_1)L_u^{S'_{u_1}\otimes\varphi_u}\right)}\right]\right\} \\
&= \sup_{C_2>0}\left\{\mathbb{E}_{W,S_{u_1}|\widetilde{z},\varphi_u}\left[C_2\left(L_u^{\bar{S}_u}-(1+C_1)L_u^{S_u}\right)\right]\right. \\
&\qquad \left. - \log \frac{\mathbb{E}_{W|\widetilde{z},\varphi_u}\left[e^{C_2L_u^{\varphi_u^+}-C_2(1+C_1)L_u^{\varphi_u^-}} + e^{C_2L_u^{\varphi_u^-}-C_2(1+C_1)L_u^{\varphi_u^+}}\right]}{2}\right\}. \qquad \text{(附录 A-73)}
\end{aligned}
$$

通过合理选择 C_1 与 C_2 的值，可保证上式右侧第二项小于 0。注意到 $e^{C_2L_u^{\varphi_u^+}-C_2(1+C_1)L_u^{\varphi_u^-}}$ 对于 $L_u^{\varphi_u^+}$ 与 $L_u^{\varphi_u^-}$ 满足联合凸性，则其最大值应取自于 $L_u^{\varphi_u^+}, L_u^{\varphi_u^-} \in [0,1]$ 的端点。若 $L_u^{\varphi_u^+} = L_u^{\varphi_u^-} = 0$，则自然有 $e^{C_2L_u^{\varphi_u^+}-C_2(1+C_1)L_u^{\varphi_u^-}} = e^{C_2L_u^{\varphi_u^-}-C_2(1+C_1)L_u^{\varphi_u^+}} = 1$。若 $L_u^{\varphi_u^+} = L_u^{\varphi_u^-} = 1$，同样

有 $e^{C_2L_u^{\varphi_u^+}-C_2(1+C_1)L_u^{\varphi_u^-}}=e^{C_2L_u^{\varphi_u^-}-C_2(1+C_1)L_u^{\varphi_u^+}}=e^{-C_2C_1}\leq 1$。否则当 $L_u^{\varphi_u^+}=0$ 且 $L_u^{\varphi_u^-}=1$（或 $L_u^{\varphi_u^+}=1$ 且 $L_u^{\varphi_u^-}=0$，可选择足够大的 C_1 值以确保 $e^{-C_2(C_1+1)}+e^{C_2}\leq 2$。求解该不等式可得 $C_1\geq -\frac{\log(2-e^{C_2})}{C_2}-1$ 且 $C_2\leq \log 2$。在此条件下，可得

$$\mathbb{E}_{W,S_{u_1}|\widetilde{z},\varphi_u}\left[C_2\left(L_u^{\bar{S}_u}-(1+C_1)L_u^{S_u}\right)\right]\leq \widetilde{I}^{\widetilde{z},\varphi_u}(W;S_{u_1}).$$

将上式代入式 (附录 A-72)，可得

$$\begin{aligned}\overline{\text{gen}}&=L-(1+C_1)L_n+C_1L_n\leq C_1L_n+\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{Z}},\Phi_u}\left[\frac{\widetilde{I}^{\widetilde{\mathbf{Z}},\Phi_u}(W;S_{u_1})}{C_2}\right]\\&=C_1L_n+\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\frac{I(W;S_{u_1}|\widetilde{\mathbf{Z}},\Phi_u)}{C_2}\leq C_1L_n+\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\frac{I(W;S_u|\widetilde{\mathbf{Z}})}{C_2}.\end{aligned}$$

通过与定理 4.6 相似的证明步骤，可得对于任意 $k\in[1,\frac{n}{m}]$ 有

$$\overline{\text{gen}}\leq C_1L_n+\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\frac{I(W;S_u|\widetilde{\mathbf{Z}})}{C_2}\leq C_1L_n+\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{I(W;S_u|\widetilde{\mathbf{Z}})}{kC_2}.$$

对于 $L_n=0$，取 $C_2\to\frac{\log 2}{2}$ 与 $C_1\to\infty$，则有 $L\leq\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{I(W;S_u|\widetilde{\mathbf{Z}})}{k\log 2}$。 ■

证明 (定理 4.12)**:** 对于任意 $t\in[1,T]$，对于 Markov 链 $Z\to(W_{T-1},\eta_tG_T+N_T)\to W_{T-1}+\eta_tG_T+N_T$ 应用数据处理不等式，可得

$$\begin{aligned}I(W_T;Z)&=I(W_{T-1}+\eta_tG_T+N_T;Z)\leq I(W_{T-1},\eta_tG_T+N_T;Z)\\&=I(W_{T-1};Z)+I(\eta_tG_T+N_T;Z|W_{T-1})\leq\cdots\leq\sum_{t=1}^{T}I(\eta_tG_t+N_t;Z|W_{t-1}).\end{aligned}$$

由于 N_t 与 Z 和 B_t 独立，故有

$$\text{Cov}_{Z,B_t,N_t}\left[\eta_tG_t+N_t\right]=\text{Cov}_{Z,B_t}\left[\eta_tG_t\right]+\text{Cov}_{N_t}[N_t]=\eta_t^2\Sigma_t+\sigma_t^2I_d.$$

在引理 附录 A.44 中取 $\Sigma=\eta_t^2\Sigma_t+\sigma_t^2I_d$，可得

$$\begin{aligned}I^{w_{t-1}}(\eta_tG_t+N_t;Z)&=H(\eta_tG_t+N_t|W_{t-1}=w)-H(\eta_tG_t+N_t|Z,W_{t-1}=w)\\&\leq H(\eta_tG_t+N_t|W_{t-1}=w)-H(\eta_tG_t+N_t|Z,B_t,W_{t-1}=w)\\&=H(\eta_tG_t+N_t|W_{t-1}=w)-H(N_t)\\&\leq\frac{d}{2}\log(2\pi e)+\frac{1}{2}\log\left|\eta_t^2\Sigma_t+\sigma_t^2I_d\right|-\frac{d}{2}\log(2\pi e\sigma_t^2)=\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\Sigma_t+I_d\right|.\end{aligned}$$

综合上述结果可得

$$I(W_T;Z) \le \sum_{t=1}^{T} I(\eta_t G_t + N_t; Z|W_{t-1}) = \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}}\left[I^{W_{t-1}}(\eta_t G_t + N_t; Z|W_{t-1})\right]$$
$$= \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}}\left[\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\Sigma_t + I_d\right|\right] \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d\right|,$$

其中最后一步可通过对对数行列式函数应用 Jensen 不等式得到。 ■

证明 (定理 4.14)**:** 根据期望泛化误差的定义，可得

$$|\overline{\text{gen}}| = \left|\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,\widetilde{Z}_u,S_u}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right]\right| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_u,L_u}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right]\right|. \tag{附录 A-74}$$

对于任意 $u\in\mathsf{P}_n^m$，有 $L_u^{\bar{S}_u} - L_u^{S_u} \in [-1,1]$。因此，$L_u^{\bar{S}_u} - L_u^{S_u}$ 满足 1-次高斯性。在引理 附录 A.7 中取 $f(L_u,S_u) = L_u^{\bar{S}_u} - L_u^{S_u}$，可得

$$\left|\mathbb{E}_{S_u,L_u}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right] - \mathbb{E}_{S_u',L_u}\left[L_u^{\bar{S}_u'} - L_u^{S_u'}\right]\right| \le \sqrt{2I(L_u,S_u)}.$$

易证 $\mathbb{E}_{S_u',L_u}\left[L_u^{\bar{S}_u'} - L_u^{S_u'}\right] = 0$。代入式 (附录 A-74)，即得 $|\overline{\text{gen}}| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\sqrt{2I(L_u,S_u)}$。 ■

证明 (定理 4.15)**:** 根据期望泛化误差的定义，可得

$$|\overline{\text{gen}}| = \left|\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,\widetilde{Z}_u,S_u}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right]\right| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_u,L_u}\left[L_u^{\bar{S}_u} - L_u^{S_u}\right]\right|$$
$$= \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\left(L_u^{\Phi_u^+} - L_u^{\Phi_u^-}\right)\right]\right| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\right]\right|. \tag{附录 A-75}$$

对于任意 $u\in\mathsf{P}_n^m$，有 $\Delta_u^{\Phi_u} \in [-1,1]$。因此，$(-1)^{S_{u_1}'}\Delta_u^{\Phi_u}$ 满足 1-次高斯性。通过在引理 附录 A.7 中取 $f(S_{u_1},\Delta_u^{\Phi_u}) = (-1)^{S_{u_1}}\Delta_u^{\Phi_u}$，可得

$$\left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\right] - \mathbb{E}_{S_u',L_u,\Phi_u}\left[(-1)^{S_{u_1}'}\Delta_u^{\Phi_u}\right]\right| \le \sqrt{2I(\Delta_u^{\Phi_u};S_{u_1})}.$$

注意到 $\mathbb{E}_{S_u',L_u,\Phi_u}\left[(-1)^{S_{u_1}'}\Delta_u^{\Phi_u}\right] = 0$，则将上式代入式 (附录 A-75)，可得

$$|\overline{\text{gen}}| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\right]\right| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\sqrt{2I(\Delta_u^{\Phi_u};S_{u_1})}.$$ ■

证明 (定理 4.16)**:** 根据条件 $\ell(\cdot,\cdot)\in\{0,1\}$ 以及 $L_n = 0$，有

$$L = \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\frac{\mathbb{E}_{L_u,\Phi_u|S_{u_1}=0}\left[L_u^{\Phi_u^+}\right] + \mathbb{E}_{L_u,\Phi_u|S_{u_1}=1}\left[L_u^{\Phi_u^+}\right]}{2}$$

$$= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{P\left(\Delta_u^{\Phi_u} = 1 | S_{u_1} = 0\right) + P\left(\Delta_u^{\Phi_u} = -1 | S_{u_1} = 1\right)}{2}. \tag{附录 A-76}$$

注意到训练损失项 $L_u^{S_u}$（或测试损失项 $L_u^{\bar{S}_u}$）的分布应与 S_{u_1} 的取值无关。因此，给定 S_{u_1} 时，$L_u^{\Phi_u^+}$ 与 $L_u^{\Phi_u^-}$ 的分布应互相对称，即有 $P_{L_u^{\Phi_u^+}|S_{u_1}=0} = P_{L_u^{\Phi_u^-}|S_{u_1}=1}$ 且 $P_{L_u^{\Phi_u^+}|S_{u_1}=1} = P_{L_u^{\Phi_u^-}|S_{u_1}=0}$。随之可得 $P(\Delta_u^{\Phi_u} = 1|S_{u_1} = 0) = P(\Delta_u^{\Phi_u} = -1|S_{u_1} = 1)$，$P(\Delta_u^{\Phi_u} = 0|S_{u_1} = 0) = P(\Delta_u^{\Phi_u} = 0|S_{u_1} = 1)$ 且 $P(\Delta_u^{\Phi_u} = 1|S_{u_1} = 1) = P(\Delta_u^{\Phi_u} = -1|S_{u_1} = 0) = 0$。

设 $\alpha_u = P(\Delta_u^{\Phi_u} = 1|S_{u_1} = 0)$，则有 $P(\Delta_u^{\Phi_u} = 0|S_{u_1} = 0) = 1 - \alpha_u$ 且

$$\begin{aligned} I(\Delta_u^{\Phi_u}; S_{u_1}) &= H(\Delta_u^{\Phi_u}) - H(\Delta_u^{\Phi_u}|S_{u_1}) = H\left(\frac{\alpha_u}{2}, 1-\alpha_u, \frac{\alpha_u}{2}\right) - H(\alpha_u, 1-\alpha_u) \\ &= -\alpha_u \log\left(\frac{\alpha_u}{2}\right) + \alpha_u \log(\alpha_u) = \alpha_u \log 2. \end{aligned}$$

将上式代入式 (附录 A-76)，可得

$$|\overline{\text{gen}}| = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \alpha_u = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{I(\Delta_u^{\Phi_u}; S_{u_1})}{\log 2}.$$

假设 $L_n = 0$，则有 $P(L_u^{\Phi_u^+} = 1, L_u^{\Phi_u^-} = 1) = 0$。因此，存在从 $\Delta_u^{\Phi_u}$ 到 $L_u^{\Phi_u}$ 的双射：$\Delta_u^{\Phi_u} = 0 \leftrightarrow L_u^{\Phi_u} = \{0, 0\}$，$\Delta_u^{\Phi_u} = 1 \leftrightarrow L_u^{\Phi_u} = \{0, 1\}$ 且 $\Delta_u^{\Phi_u} = -1 \leftrightarrow L_u^{\Phi_u} = \{1, 0\}$。根据数据处理不等式，可得 $I(\Delta_u^{\Phi_u}; S_{u_1}) = I(L_u^{\Phi_u}; S_{u_1})$。定理得证。 ■

证明 (定理 4.17)**:** 根据式 (附录 A-75)，可得

$$|\overline{\text{gen}}| \le \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left| \mathbb{E}_{S_u, L_u, \Phi_u}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] \right| \le \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{\mathbf{Z}}} \left| \mathbb{E}_{S_u, L_u, \Phi_u | \widetilde{\mathbf{Z}}}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] \right|. \tag{附录 A-77}$$

设 S' 为 S 的一个独立拷贝，满足 $S' \perp\!\!\!\perp W|\widetilde{\mathbf{Z}} = \widetilde{z}$。对于任意 $u \in \mathsf{P}_n^m$，在引理 附录 A.9 中取 $P = P_{S_{u_1}, L_u, \Phi_u | \widetilde{z}}$，$Q = P_{L_u, \Phi_u | \widetilde{z}} P_{S_{u_1}}$ 与 $f(S_{u_1}, \Delta_u^{\Phi_u}) = (-1)^{S_{u_1}} \Delta_u^{\Phi_u}$，可得

$$\widetilde{F}(\Delta_u^{\Phi_u}; S_{u_1}) \ge \sup_{t \in \mathbb{R}} \left\{ \mathbb{E}_{S_u, L_u, \Phi_u | \widetilde{\mathbf{Z}}}\left[t(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] - \log \mathbb{E}_{S'_u, L_u, \Phi_u | \widetilde{\mathbf{Z}}}\left[e^{t(-1)^{S'_{u_1}} \Delta_u^{\Phi_u}}\right] \right\}. \tag{附录 A-78}$$

注意到 $(-1)^{S'_{u_1}} \Delta_u^{\Phi_u} \in [-1, 1]$，根据次高斯性的定义，有 $\mathbb{E}_{S'_u, L_u, \Phi_u | \widetilde{\mathbf{Z}}}\left[e^{t(-1)^{S'_{u_1}} \Delta_u^{\Phi_u}}\right] \le e^{\frac{t^2}{2}}$。将其代入式 (附录 A-78)，可得

$$\left| \mathbb{E}_{S_u, L_u, \Phi_u | \widetilde{\mathbf{Z}}}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] \right| \le \sqrt{2\widetilde{I}(\Delta_u^{\Phi_u}; S_{u_1})}.$$

将上式代入式 (附录 A-77)，可得

$$|\overline{\text{gen}}| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{z}}}\left|\mathbb{E}_{S_u,L_u,\Phi_u|\widetilde{\mathbf{z}}}\left[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\right]\right| \le \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{\mathbf{z}}}\sqrt{2\widetilde{F}(\Delta_u^{\Phi_u};S_{u_1})}. \qquad \blacksquare$$

证明 (定理 4.19)**:** 最小值中第一项可通过与定理 4.10 相似的方法证明，将式 (附录 A-73) 中的互信息替换为 $I(L_u^{\Phi_u};S_{u_1})$ 即得证。对于第二项，注意到

$$\begin{aligned}
L-(1+C_1)L_n &= \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{S_u,L_u}\left[L_u^{\bar{S}_u}-(1+C_1)L_u^{S_u}\right]\\
&= \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{S_u,L_u}\left[\left(1+\frac{C_1}{2}\right)\left(L_u^{\bar{S}_u}-L_u^{S_u}\right)-\frac{C_1}{2}L_u^{\bar{S}_u}-\frac{C_1}{2}L_u^{S_u}\right]\\
&= \frac{1}{2|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left(\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[(C_1+2)(-1)^{S_{u_1}}L_u^{\Phi_u^+}-C_1L_u^{\Phi_u^+}\right]\right.\\
&\qquad \left.+\,\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[-(C_1+2)(-1)^{S_{u_1}}L_u^{\Phi_u^-}-C_1L_u^{\Phi_u^-}\right]\right). \qquad \text{(附录 A-79)}
\end{aligned}$$

易证 $\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[(-1)^{S_{u_1}}L_u^{\Phi_u^+}\right]=-\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[(-1)^{S_{u_1}}L_u^{\Phi_u^-}\right]$。此外，注意到 $P_{L_u^{\Phi_u^+}}=P_{L_u^{\Phi_u^-}}$，可得 $\mathbb{E}_{L_u^{\Phi_u^+}}\left[L_u^{\Phi_u^+}\right]=\mathbb{E}_{L_u^{\Phi_u^-}}\left[L_u^{\Phi_u^-}\right]$。将其代入式 (附录 A-79)，则有

$$L-(1+C_1)L_n = \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[(C_1+2)(-1)^{S_{u_1}}L_u^{\Phi_u^+}-C_1L_u^{\Phi_u^+}\right]. \qquad \text{(附录 A-80)}$$

对于任意 $u\in\mathsf{P}_n^m$，在引理 附录 A.9 中取 $P=P_{L_u^{\Phi_u^+},S_{u_1}}$，$Q=P_{L_u^{\Phi_u^+}}P_{S_{u_1}}$ 与 $f(L_u^{\Phi_u^+},S_{u_1})=C_2(C_1+2)(-1)^{S_{u_1}}L_u^{\Phi_u^+}-C_2C_1L_u^{\Phi_u^+}$，可得

$$\begin{aligned}
I(L_u^{\Phi_u^+};S_{u_1}) \ge \sup_{C_2>0}\Bigg\{&\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[C_2(C_1+2)(-1)^{S_{u_1}}L_u^{\Phi_u^+}-C_2C_1L_u^{\Phi_u^+}\right]\\
&-\log\frac{\mathbb{E}_{L_u,\Phi_u}\left[e^{-2C_2(C_1+1)L_u^{\Phi_u^+}}+e^{2C_2L_u^{\Phi_u^+}}\right]}{2}\Bigg\}. \qquad \text{(附录 A-81)}
\end{aligned}$$

通过选择合适的 C_1 与 C_2 值，即可保证上式右侧的第二项小于 0。注意到 $e^{-2C_2(C_1+1)L_u^{\Phi_u^+}}$ 与 $e^{2C_2L_u^{\Phi_u^+}}$ 均为关于 $L_u^{\Phi_u^+}$ 的凸函数，则该项最大值应取自于 $L_u^{\Phi_u^+}\in[0,1]$ 的端点。若 $L_u^{\Phi_u^+}=0$，则自然有 $e^{-2C_2(C_1+1)L_u^{\Phi_u^+}}+e^{2C_2L_u^{\Phi_u^+}}=2$。否则若 $L_u^{\Phi_u^+}=1$，则可选择足够大的 C_1 以确保 $e^{-2C_2(C_1+1)}+e^{2C_2}\le 2$。求解该不等式可得 $C_1\ge-\frac{\log(2-e^{2C_2})}{2C_2}-1$ 且 $C_2\le\frac{\log 2}{2}$。在此条件下，对于任意 $u\in\mathsf{P}_n^m$，有

$$\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[C_2(C_1+2)(-1)^{S_{u_1}}L_u^{\Phi_u^+}-C_2C_1L_u^{\Phi_u^+}\right]\le I(L_u^{\Phi_u^+};S_{u_1}).$$

将上述不等式代入式 (附录 A-80)，可得

$$\overline{\text{gen}} = L - (1+C_1)L_n + C_1 L_n \le C_1 L_n + \sum_{u\in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| C_2}.$$

当 $L_n = 0$ 时，取 $C_2 \to \frac{\log 2}{2}$ 与 $C_1 \to \infty$，即可得 $L \le \sum_{u\in \mathsf{P}_n^m} \frac{2I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| \log 2}$。 ■

证明 (定理 4.20)**:** 根据损失方差的定义，有

$$\begin{aligned}
V(\gamma) &= \mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left(\ell(W,\widetilde{Z}_u^{S_u}) - (1+\gamma)L_{\widetilde{\mathbf{Z}}_S}(W)\right)^2\right] \\
&= \mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left(\ell^2(W,\widetilde{Z}_u^{S_u}) - 2(1+\gamma)\ell(W,\widetilde{Z}_u^{S_u})L_{\widetilde{\mathbf{Z}}_S}(W) + (1+\gamma)^2 L^2_{\widetilde{\mathbf{Z}}_S}(W)\right)\right] \\
&= \mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\ell(W,\widetilde{Z}_u^{S_u})\right] - 2(1+\gamma)\mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L^2_{\widetilde{\mathbf{Z}}_S}(W)\right] + (1+\gamma)^2\mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L^2_{\widetilde{\mathbf{Z}}_S}(W)\right] \\
&= L_n - (1-\gamma^2)\mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L^2_{\widetilde{\mathbf{Z}}_S}(W)\right].
\end{aligned}$$

由于 $\ell(\cdot,\cdot) \in \{0,1\}$，可得 $L_{\widetilde{\mathbf{Z}}_S}(W) \in [0,1]$，$L^2_{\widetilde{\mathbf{Z}}_S}(W) \le L_{\widetilde{\mathbf{Z}}_S}(W)$ 且

$$\begin{aligned}
\overline{\text{gen}} - C_1 V(\gamma) &= \overline{\text{gen}} - C_1 L_n + C_1(1-\gamma^2)\mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L^2_{\widetilde{\mathbf{Z}}_S}(W)\right] \\
&\le \overline{\text{gen}} - C_1 L_n + C_1(1-\gamma^2)\mathbb{E}_{W,\widetilde{\mathbf{Z}},S}\left[L_{\widetilde{\mathbf{Z}}_S}(W)\right] = \overline{\text{gen}} - C_1\gamma^2 L_n. \quad \text{(附录 A-82)}
\end{aligned}$$

在定理 4.19 中取 $C_1 = C_1\gamma^2$ 与 $C_2 = C_2$，可得 $\overline{\text{gen}} - C_1\gamma^2 L_n \le \sum_{u\in\mathsf{P}_n^m} \frac{I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| C_2}$。结合式 (附录 A-82) 则定理得证。 ■

证明 (定理 4.21)**:** 在式 (附录 A-70) 中，证明了

$$d\left(L_n \middle\| \frac{L_n + L}{2}\right) \le \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{L_u^{\Phi_u},S_{u_1}}\left[d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right)\right]. \quad \text{(附录 A-83)}$$

对于任意 $u \in \mathsf{P}_n^m$，在引理 附录 A.9 中取 $P = P_{L_u^{\Phi_u},S_{u_1}}$，$Q = P_{L_u^{\Phi_u}}P_{S_{u_1}}$ 与 $f(L_u^{\Phi_u}, S_{u_1}) = d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right)$，可得

$$I(L_u^{\Phi_u}, S_{u_1}) \ge \mathbb{E}_{L_u^{\Phi_u},S_{u_1}}\left[d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right)\right] - \log\mathbb{E}_{L_u^{\Phi_u},S'_{u_1}}\left[\exp\left(d_\gamma\left(L_u^{S'_{u_1}\otimes\Phi_u}\middle\|\frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right)\right)\right]. \quad \text{(附录 A-84)}$$

注意到 $\mathbb{E}_{S'_{u_1}}\left[L_u^{S'_{u_1}\otimes\Phi_u}\right] = \frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2} \in [0,1]$，通过应用引理 附录 A.12，可得对于任意 $\gamma \in \mathbb{R}$，

有 $\mathbb{E}_{L_u^{\Phi_u},S'_{u_1}}\left[\exp\left(d_\gamma\left(L_u^{S'_{u_1}\otimes\Phi_u}\middle\|\frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right)\right)\right]\leq 1$。将其代入式 (附录 A-84)，可得

$$\mathbb{E}_{L_u^{\Phi_u},S_{u_1}}\left[d_\gamma\left(L_u^{S_u}\middle\|\frac{L_u^{\Phi_u^+}+L_u^{\Phi_u^-}}{2}\right)\right]\leq I(L_u^{\Phi_u},S_{u_1}).$$

将上式代入式 (附录 A-83)，则定理得证。 ■

A.5 第 5 章定理补充证明

引理 附录 A.45 设 P 与 Q 为定义在相同空间上的概率分布，$X\sim P$ 且 $X'\sim Q$。若 $f(X)$ 对于 X 满足 σ-次高斯性且以下期望存在，则

$$\left|\mathbb{E}_{X'}[f(X')]-\mathbb{E}_X[f(X)]\right|\leq\sqrt{2\sigma^2D(Q\,\|\,P)}.$$

证明: 设 $\lambda\in\mathbb{R}$ 为任意非 0 常数，则根据 $f(X)$ 的次高斯性可得

$$\log\mathbb{E}_X\left[e^{\lambda f(X)}\right]-\lambda\mathbb{E}_X[f(X)]\leq\frac{\lambda^2\sigma^2}{2}.$$

在引理 附录 A.9 中取 $X=\lambda f(X)$，则有

$$\begin{aligned}D(Q\,\|\,P)&\geq\sup_\lambda\left\{\mathbb{E}_{X'}[\lambda f(X')]-\log\mathbb{E}_X\left[e^{\lambda f(X)}\right]\right\}\geq\sup_\lambda\left\{\mathbb{E}_{X'}[\lambda f(X')]-\lambda\mathbb{E}_X[f(X)]-\frac{\lambda^2\sigma^2}{2}\right\}\\&=\frac{1}{2\sigma^2}\Big(\mathbb{E}_{X'}[f(X')]-\mathbb{E}_X[f(X)]\Big)^2,\end{aligned}$$

其中最大值取自于 $\lambda=\frac{1}{\sigma^2}(\mathbb{E}_{X'}[f(X')]-\mathbb{E}_X[f(X)])$。 ■

证明 (定理 5.7)**:** 简便起见，对于任意随机变量 X，将 $P_X(X)$ 简写为 P_X。

$$\begin{aligned}D(P_{Y|X,D}\,\|\,Q_{Y|X})&=\mathbb{E}_{D,X,Y}\left[\log\frac{P_{Y|X,D}}{Q_{Y|X}}\right]=\mathbb{E}_{D,X,Y}\left[\log\frac{P_{Y|X,D}}{P_{Y|X}}\cdot\frac{P_{Y|X}}{Q_{Y|X}}\right]\\&=\mathbb{E}_{D,X,Y}\left[\log\frac{P_{Y,D|X}}{P_{Y|X}P_{D|X}}\right]+\mathbb{E}_{X,Y}\left[\log\frac{P_{Y|X}}{Q_{Y|X}}\right]\\&=I(Y;D|X)+D(P_{Y|X}\,\|\,Q_{Y|X})\geq I(Y;D|X).\end{aligned}$$

最后一步可由 KL 散度的正值性得到，当且仅当 $Q_{Y|X}=P_{Y|X}$ 时上述不等式取等号。 ■

证明 (定理 5.8)**:** 对于任意 $D\in D_s$，设 $\bar{D}$ 为 D 的一个独立拷贝。由 $\ell(\cdot,\cdot)\in[0,M]$ 可知 $L_{\bar{D}}(W)$ 满足 $\frac{M}{2}$-次高斯性。在引理 附录 A.7 中取 $X=W$，$Y=D$ 与 $f(W,D)=L_D(W)$ 可得

$$|\mathbb{E}_{W,D}[L_D(W)] - \mathbb{E}_W[L(W)]| = |\mathbb{E}_{W,D}[L_D(W)] - \mathbb{E}_{W,\bar{D}}[L_{\bar{D}}(W)]| \leq \sqrt{\frac{M^2}{2} I(W,D)}.$$

将上述不等式对每一个源域求和，则可得

$$|\mathbb{E}_{W,D_s}[L_s(W)] - \mathbb{E}_W[L(W)]| \leq \frac{1}{m}\sum_{i=1}^{m}\left|\mathbb{E}_{W,D_i}[L_{D_i}(W)] - \mathbb{E}_W[L(W)]\right| \leq \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{M^2}{2} I(W,D_i)}.$$

类似地，对于任意 $D \in D_s$，可验证 $|L_{\bar{D}}(W) - L(W)| \in [0, M]$，从而满足 $\frac{M}{2}$-次高斯性。在引理 附录 A.7 中取 $X = W$，$Y = D$ 与 $f(W,D) = |L_D(W) - L(W)|$ 可得

$$\mathbb{E}_{W,D}|L_D(W) - L(W)| - \mathbb{E}_{W,\bar{D}}|L_{\bar{D}}(W) - L(W)| \leq \sqrt{\frac{M^2}{2} I(W,D)}.$$

将上述不等式对每一个源域求和，则可得

$$\begin{aligned}\mathbb{E}_{W,D_s}|L_s(W) - L(W)| &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{W,D_i}|L_{D_i}(W) - L(W)| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{M^2}{2} I(W,D_i)} + \mathbb{E}_{W,\bar{D}}|L_{\bar{D}}(W) - L(W)|.\end{aligned}$$

应用 Markov 不等式，有

$$\mathbb{P}(|L_s(W) - L(W)| \geq \varepsilon) \leq \frac{M}{m\varepsilon\sqrt{2}}\sum_{i=1}^{m}\sqrt{I(W,D_i)} + \frac{1}{\varepsilon}\mathbb{E}_{W,\bar{D}}|L_{\bar{D}}(W) - L(W)|. \qquad \blacksquare$$

证明 (定理 5.12)**:** 对于任意 $D_i \in D_s$，在引理 附录 A.10 中取 $P = P_{W|D_i=d}$，$Q = P_W$ 与 $f(w) = L_{D_i}(w)$ 可得

$$\begin{aligned}|\mathbb{E}_{W,D_s}[L_s(W)] - \mathbb{E}_W[L(W)]| &\leq \frac{1}{m}\mathbb{E}_{D_s}\left[\sum_{i=1}^{m}|\mathbb{E}_{W|D_i}[L_{D_i}(W)] - \mathbb{E}_W[L_{D_i}(W)]|\right] \\ &\leq \frac{1}{m}\mathbb{E}_{D_s}\left[\sum_{i=1}^{m}\beta'\mathbb{W}(P_{W|D_i}, P_W)\right] = \frac{\beta'}{m}\sum_{i=1}^{m}\mathbb{E}_{D_i}[\mathbb{W}(P_{W|D_i}, P_W)].\blacksquare\end{aligned}$$

证明 (定理 5.13)**:** 对于任意 $d \in \mathcal{D}$，在引理 附录 A.45 中取 $P = P_Z$，$Q = P_{Z|D=d}$ 与 $f(Z) = \ell(f_w(X), Y)$，则有

$$|L_d(w) - L(w)| = \left|\mathbb{E}_{Z|D=d}[\ell(f_w(X), Y)] - \mathbb{E}_Z[\ell(f_w(X), Y]\right| \leq \sqrt{2\sigma^2 D(P_{Z|D=d} \,\|\, P_Z)}.$$

将上式对 $D \sim v$ 取期望，则有

$$\mathbb{E}_D|L_d(w) - L(w)| \leq \mathbb{E}_D\sqrt{2\sigma^2 D(P_{Z|D} \,\|\, P_Z)} \leq \sqrt{2\sigma^2\mathbb{E}_D[D(P_{Z|D} \,\|\, P_Z)]} = \sigma\sqrt{2I(Z;D)}.$$

将上式对每一个目标域求和，则可得

$$\mathbb{E}_{D_t}|L_t(w)-L(w)| \le \frac{1}{m'}\sum_{k=1}^{m'}\mathbb{E}_{D_k}\left|L_{D_k}(w)-L(W)\right| \le \frac{1}{m'}\sum_{k=1}^{m'}\sigma\sqrt{2I(Z;D_k)} = \sigma\sqrt{2I(Z;D)}.$$

应用 Markov 不等式，即可得

$$\mathbb{P}(|L_t(w)-L(w)| \ge \varepsilon) \le \frac{\sigma}{\varepsilon}\sqrt{2I(Z;D)}.$$ ■

证明 (定理 5.14)**:** 对于任意领域 $d \in \mathcal{D}$ 与分类器 ψ and $f^* : \mathcal{R} \mapsto \mathcal{Y}$，定义

$$L_d(\psi,f^*) = \mathbb{E}_{R|D=d}[\ell(f_\psi(R),f^*(R))], \qquad L(\psi,f^*) = \mathbb{E}_R[\ell(f_\psi(R),f^*(R))].$$

在引理 附录 A.45 中取 $P = P_R$，$Q = P_{R|D=d}$ 与 $f(R) = \ell(f_\psi(R),f^*(R))$，则有

$$|L_d(\psi,f^*) - L(\psi,f^*)| \le \sqrt{2\sigma^2 D(P_{R|D=d}\,\|\,P_R)}.$$

将上式对 $D \sim \nu$ 取期望，则有

$$\mathbb{E}_D|L_D(\psi,f^*) - L(\psi,f^*)| \le \mathbb{E}_D\sqrt{2\sigma^2 D(P_{R|D}\,\|\,P_R)} \le \sqrt{2\sigma^2 I(R;D)}.$$

若假设 5.4 成立，则根据损失函数 $\ell(\cdot,\cdot)$ 的对称性与三角不等式，可得

$$L_d(\psi,f^*) = \mathbb{E}_{R|D=d}[\ell(f_\psi(R),f^*(R))] \le \mathbb{E}_{R,Y|D=d}[\ell(f_\psi(R),Y) + \ell(f^*(R),Y)] = L_d(\psi) + L_d(f^*).$$

同理可得 $L(\psi,f^*) \le L(\psi) + L(f^*)$。类似地，可证明

$$L_d(\psi,f^*) = \mathbb{E}_{R|D=d}[\ell(f_\psi(R),f^*(R))] \ge \mathbb{E}_{R,Y|D=d}[\ell(f_\psi(R),Y) - \ell(f^*(R),Y)] = L_d(\psi) - L_d(f^*).$$

同理有 $L(\psi,f^*) \ge L(\psi) - L(f^*)$。综合上述结果，有

$$\begin{aligned} L_d(\psi) - L(\psi) &\le L_d(\psi,f^*) + L_d(f^*) - L(\psi,f^*) + L(f^*), \\ L(\psi) - L_d(\psi) &\le L(\psi,f^*) + L(f^*) - L_d(\psi,f^*) + L_d(f^*). \end{aligned}$$

结合上述不等式，可得

$$|L_d(\psi) - L(\psi)| \le |L_d(\psi,f^*) - L(\psi,f^*)| + L_d(f^*) + L(f^*).$$

将上式对 $D \sim \nu$ 取期望，则有

$$\begin{aligned} \mathbb{E}_D|L_D(\psi) - L(\psi)| &\le \mathbb{E}_D|L_D(\psi,f^*) - L(\psi,f^*)| + \mathbb{E}_D[L_D(f^*) + L(f^*)] \\ &\le \sqrt{2\sigma^2 I(R;D)} + 2L(f^*). \end{aligned}$$

类似地，将上式对每一个目标域求和，可得

$$\mathbb{E}_{D_t}|L_t(\psi) - L(\psi)| \leq \sqrt{2\sigma^2 I(R;D)} + 2L(f^*).$$

最后，应用 Markov 不等式，可得

$$\mathbb{P}(L_t(\psi) - L(\psi) \geq \varepsilon) \leq \frac{\sigma}{\varepsilon}\sqrt{2I(R;D)} + \frac{2}{\varepsilon}L(f^*).$$

在上式中取最小值 $\min_{f^*}[L(f^*)]$，定理得证。 ■

证明 (定理 5.16)**:** 绝对连续条件 $P_{R,D} \ll P_{R_i,D_i}$ 与 $P_{R_i,D_i} \ll P_{R,D}$ 隐含着存在 $B > 1$，使得对于任意 $r \in \mathcal{R}$ 与 $d \in \mathcal{D}$，有 $\frac{P_{R,D}(r,d)}{P_{R_i,D_i}(r,d)} \in [\frac{1}{B}, B]$。因此，$\log \frac{P_{R,D}}{P_{R_i,D_i}} \in [-\log(B), \log(B)]$ 且满足 $\log(B)$-次高斯性。

在引理 附录 A.7 中取 $X = W$，$Y = D_i$ 与 $f(W,D) = \mathbb{E}_{X|D}[\log \frac{P_{R,D}}{P_{R_i,D_i}}]$，可得

$$\begin{aligned}
\left|\mathbb{E}_{W,D_i,R_i}\left[\log \frac{P_{R,D}(R_i,D_i)}{P_{R_i,D_i}(R_i,D_i)}\right] - \mathbb{E}_{W,D,R}\left[\log \frac{P_{R,D}(R,D)}{P_{R_i,D_i}(R,D)}\right]\right| &\leq \sqrt{2\log^2(B) I(W;D_i)} \\
\left|D(P_{R,D} \,\|\, P_{R_i,D_i}) + D(P_{R_i,D_i} \,\|\, P_{R,D})\right| &\leq \sqrt{2\log^2(B) I(W;D_i)} \\
D_s(P_{R,D} \,\|\, P_{R_i,D_i}) &\leq \log(B)\sqrt{2I(W;D_i)}.
\end{aligned}$$ ■

证明 (定理 5.10)**:** 注意到 Markov 链 $D_i \to (W_{T-1}, G_T) \to W_{T-1} + G_T$，则根据数据处理不等式有

$$I(W_T;D_i) = I(W_{T-1} + G_T; D_i) \leq I(W_{T-1}, G_T; D_i) = I(W_{T-1};D_i) + I(G_T;D_i|W_{T-1}).$$

其中最后一步可通过条件互信息的链式法则得到。重复应用以上推导，可得

$$I(W_T;D_i) \leq I(W_{T-1};D_i) + I(G_T;D_i|W_{T-1}) \leq \cdots \leq \sum_{t=1}^{T} I(G_t;D_i|W_{t-1}).$$ ■

定理 附录 A.46 (定理 5.18 的正式表述) 设 n 为数据点维度，b 为数据点个数，则

- 若 $n > b + 1$，则对于任意数据点集 $s = \{x_i\}_{i=1}^b$，存在无限多个领域 $d_1, d_2, \cdots$，使得 $p(S = s|D = d_1) = p(S = s|D = d_2) = \cdots$。
- 若 $n > 2b + 1$，则对于任意两组数据点集 $s_1 = \{x_i^1\}_{i=1}^b$ 与 $s_2 = \{x_i^2\}_{i=1}^b$，存在无限多个领域 $d_1, d_2, \cdots$，使得对于任意 $j \in [1, \infty)$，有 $p(S = s_1|D = d_j) = p(S = s_2|D = d_j)$。

证明: 不失一般性地，假设数据分布 $p(X|D)$ 为均值为 0 的高斯分布，即

$$p(x|D = d) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_d|}} \exp\left(-\frac{1}{2}x^\top \Sigma_d x\right),$$

其中 Σ_d 为领域 d 对应的协方差矩阵。设 $X \in \mathbb{R}^{n\times b}$ 为数据点集 S 对应的数据矩阵，其中

X 的第 i 列即为数据点 x_i，则有

$$p(S=s|D=d)=\frac{1}{\sqrt{(2\pi)^{bn}|\Sigma_d|^b}}\exp\left(-\frac{1}{2}\mathrm{tr}(X^\top\Sigma_d X)\right).$$

由于矩阵 X 的秩至多为 b，可通过特征值分解将 Σ_d 拆分为 $\Sigma_d=\Sigma_d^1+\Sigma_d^2$，其中 $\mathrm{rank}(\Sigma_d^1)=b$ 且 $\mathrm{rank}(\Sigma_d^2)=n-b\geq 2$。设 Σ_d^1 的特征空间包含 X 的列空间，则有

$$\mathrm{tr}(X^\top\Sigma_d^1 X)=\mathrm{tr}(X^\top\Sigma_d X),\qquad \mathrm{tr}(X^\top\Sigma_d^2 X)=0.$$

因此，可任意修改矩阵 Σ_d^2 的特征空间。只要其与 Σ_d^1 的特征空间正交，便不会改变 $\mathrm{tr}(X^\top\Sigma_d X)$ 的值。定理的第一部分得证。

对于第二部分，可得类似拆分 $\Sigma_d=\Sigma_d^1+\Sigma_d^2$，其中 $\mathrm{rank}(\Sigma_d^1)=2b+1$ 且 $\mathrm{rank}(\Sigma_d^2)=n-2b-1\geq 1$。设 Σ_d^1 的特征空间包含 X_1 与 X_2 的列空间，则有

$$\mathrm{tr}(X_1^\top\Sigma_d^1 X_1)=\mathrm{tr}(X_1^\top\Sigma_d X_1),\quad \mathrm{tr}(X_2^\top\Sigma_d^1 X_2)=\mathrm{tr}(X_2^\top\Sigma_d X_2),\quad \mathrm{tr}(X_1^\top\Sigma_d^2 X_1)=\mathrm{tr}(X_2^\top\Sigma_d^2 X_2)=0.$$

设 $U_d^\top\Lambda_d U_d$ 为 Σ_d^1 的特征值分解，其中 $U_d\in\mathbb{R}^{(2b+1)\times n}$，$\Lambda_d=\mathrm{diag}(\lambda_1^d,\cdots,\lambda_{2b+1}^d)$。注意到对于任意 $x\in\mathbb{R}^n$，有 $x^\top\Sigma_d x=(U_d x)^\top\Lambda_d(U_d x)=\sum_{i=1}^{2b+1}(U_d x)_i^2\lambda_i$。假设 $p(S=s_1|D=d)=p(S=s_2|D=d)$，则有以下齐次线性方程组：

$$\begin{aligned}a_1^1\lambda_1+a_1^2\lambda_2+\cdots+a_1^{2b+1}\lambda_{2b+1}&=0,\\ &\cdots\\ a_b^1\lambda_1+a_b^2\lambda_2+\cdots+a_b^{2b+1}\lambda_{2b+1}&=0,\end{aligned}$$

其中 $a_i^j=(U_d x_i^1)_j^2-(U_d x_i^2)_j^2$。由于 $2b+1>b$，上述齐次线性方程组拥有无限个非零解，定理的第二部分得证。 ■

证明 (定理 5.19)**:** 由定义 $\bar{p}(x)=\frac{1}{b}\sum_{i=1}^b p_i(x)$ 与 $\bar{q}(x)=\frac{1}{b}\sum_{i=1}^b q_i(x)$，可得

$$\begin{aligned}D(\bar{P}\,\|\,\bar{Q})&=\int_{\mathcal{X}}\bar{p}(x)\log\left(\frac{\bar{p}(x)}{\bar{q}(x)}\right)\mathrm{d}x=-\int_{\mathcal{X}}\bar{p}(x)\log\left(\frac{1}{b}\sum_{i=1}^b\frac{p_i(x)}{\bar{p}(x)}\cdot\frac{q_{f(i)}(x)}{p_i(x)}\right)\mathrm{d}x\\ &\leq-\int_{\mathcal{X}}\bar{p}(x)\frac{1}{b}\sum_{i=1}^b\frac{p_i(x)}{\bar{p}(x)}\log\left(\frac{q_{f(i)}(x)}{p_i(x)}\right)\mathrm{d}x\\ &=-\frac{1}{b}\sum_{i=1}^b\int_{\mathcal{X}}p_i(x)\log\left(\frac{q_{f(i)}(x)}{p_i(x)}\right)\mathrm{d}x=\frac{1}{b}\sum_{i=1}^b D(P_i\,\|\,Q_{f(i)}),\end{aligned}$$

第二步可通过 Jensen 不等式得到。对于 Wasserstein 距离，应用引理 附录 A.10 可得

$$\mathbb{W}(\bar{P},\bar{Q}) = \sup_{f\in Lip_1}\left\{\int_{\mathcal{X}} f\mathrm{d}\bar{P} - \int_{\mathcal{X}} f\mathrm{d}\bar{Q}\right\} = \sup_{f\in Lip_1}\left\{\int_{\mathcal{X}} f\mathrm{d}\left(\frac{1}{b}\sum_{i=1}^{b} P_i\right) - \int_{\mathcal{X}} f\mathrm{d}\left(\frac{1}{b}\sum_{i=1}^{b} Q_{f(i)}\right)\right\}$$
$$\leq \frac{1}{b}\sum_{i=1}^{b}\sup_{f\in Lip_1}\left\{\int_{\mathcal{X}} f\mathrm{d}P_i - \int_{\mathcal{X}} f\mathrm{d}Q_{f(i)}\right\} = \frac{1}{b}\sum_{i=1}^{b}\mathbb{W}(P_i,Q_{f(i)}). \qquad \blacksquare$$

证明 (定理 5.20)**:** 简便起见，假设所有数据点 $\{x_i^1\}_{i=1}^b$ 与 $\{x_i^2\}_{i=1}^b$ 均互不相同。由于 P_i 与 Q_i 为方差相同的高斯分布，其 KL 散度或 Wasserstein 距离拥有解析表达

$$D(P_i \,\|\, Q_j) = \frac{(x_i^1 - x_j^2)^2}{2\sigma^2}, \qquad \mathcal{W}(P_i,Q_j) = |x_i^1 - x_j^2|.$$

假设存在 $i\in[1,b]$ 使得 $f(i)\neq i$。不失一般性地，假设 $f(i)>i$。则根据抽屉原理，存在 $j\in(i,b]$ 使得 $f(j)<f(i)$。假设 $\{x_i^1\}_{i=1}^b$ 与 $\{x_i^2\}_{i=1}^b$ 均升序排序，则有 $x_i^1<x_j^1$ 且 $x_{f(i)}^2>x_{f(j)}^2$。对于任意 $p\in\{1,2\}$，下列 3 种情形覆盖了所有 x_i^1，x_j^1，$x_{f(j)}^2$ 与 $x_{f(i)}^2$ 之间的大小关系：

- 当 $x_i^1<x_j^1<x_{f(j)}^2<x_{f(i)}^2$ 且 $p=2$ 时，有

$$\begin{aligned}&(x_i^1 - x_{f(i)}^2)^2 + (x_j^1 - x_{f(j)}^2)^2 - (x_i^1 - x_{f(j)}^2)^2 - (x_j^1 - x_{f(i)}^2)^2\\ &= (2x_i^1 - x_{f(i)}^2 - x_{f(j)}^2)(x_{f(j)}^2 - x_{f(i)}^2) - (2x_j^1 - x_{f(j)}^2 - x_{f(i)}^2)(x_{f(j)}^2 - x_{f(i)}^2)\\ &= (x_{f(j)}^2 - x_{f(i)}^2)(2x_i^1 - 2x_j^1) > 0.\end{aligned}$$

  否则当 $p=1$ 时，有

$$|x_i^1 - x_{f(i)}^2| + |x_j^1 - x_{f(j)}^2| = |x_i^1 - x_{f(j)}^2| + |x_j^1 - x_{f(i)}^2|.$$

- 当 $x_i^1<x_{f(j)}^2<x_j^1<x_{f(i)}^2$ 时，有

$$|x_i^1 - x_{f(i)}^2|^p > |x_i^1 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p.$$

- 当 $x_i^1<x_{f(j)}^2<x_{f(i)}^2<x_j^1$ 时，有

$$\begin{aligned}|x_i^1 - x_{f(i)}^2|^p + |x_j^1 - x_{f(j)}^2|^p &\geq |x_i^1 - x_{f(j)}^2|^p + |x_{f(i)}^2 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p + |x_{f(i)}^2 - x_{f(j)}^2|^p\\ &> |x_i^1 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p.\end{aligned}$$

综上所述，对于所有可能情况，可得 $|x_i^1 - x_{f(i)}^2|^p + |x_j^1 - x_{f(j)}^2|^p \geq |x_i^1 - x_{f(j)}^2|^p + |x_j^1 - x_{f(i)}^2|^p$。这意味着若取 $f'(i)=f(j)$，$f'(j)=f(i)$ 且对于其他 $k\notin\{i,j\}$ 取 $f'(k)=f(k)$，则 f' 相较于 f 更能够最小化 $D(\bar{P}\,\|\,\bar{Q})$ 或 $\mathcal{W}(\bar{P},\bar{Q})$。由于最小值显然存在，定理得证。 $\blacksquare$

攻读学位期间取得的研究成果

[1] **Dong Y**, Gong T, Chen H, Song S, Zhang W, Li C. How does distribution matching help domain generalization: An information-theoretic analysis[J]. IEEE Transactions on Information Theory, 2025 Mar;71(3):2028-53.（CCF-A 类期刊）

[2] **Dong Y**, Gong T, Yu S, Li C. Optimal randomized approximations for matrix-based Rényi's entropy[J]. IEEE Transactions on Information Theory, 2023 Mar 21;69(7):4218-34.（CCF-A 类期刊）

[3] **Dong Y**, Gong T, Chen H, He Z, Li M, Song S, Li C. Towards generalization beyond pointwise learning: A unified information-theoretic perspective[C].//International Conference on Machine Learning, 2024 Jul 8 (pp. 11311-11345).（CCF-A 类会议）

[4] **Dong Y**, Gong T, Chen H, Yu S, Li C. Rethinking information-theoretic generalization: Loss entropy induced PAC bounds[C].//International Conference on Learning Representations, 2024 Jan 1.（清华 A 类会议）

[5] **Dong Y**, Gong T, Chen H, Li C. Understanding the generalization ability of deep learning algorithms: A kernelized Rényi's entropy perspective[C].//International Joint Conference on Artificial Intelligence, 2023 Aug 19 (pp. 3642-3650).（CCF-A 类会议）

[6] **Dong Y**, Gong T, Yu S, Chen H, Li C. Robust and fast measure of information via low-rank representation[C].//AAAI Conference on Artificial Intelligence, 2023 Jun 26 (Vol. 37, No. 6, pp. 7450-7458).（CCF-A 类会议）

[7] **Dong Y**, Gong T, Chen H, Li C. Efficient approximations for matrix-based Rényi's entropy on sequential data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023 Sep 19.（中科院一区 Top）

[8] Gong T*, **Dong Y***, Yu S, Dong B. Computationally efficient approximations for matrix-based Rényi's entropy[J]. IEEE Transactions on Signal Processing, 2022;70:6170-84.（学生一作，中科院二区 Top）

[9] Gong T, **Dong Y**, Chen H, Feng W, Dong B, Li C. Regularized modal regression on Markov-dependent observations: A theoretical assessment[C].//AAAI Conference on Artificial Intelligence, 2022 Jun 28 (Vol. 36, No. 6, pp. 6721-6728).（学生一作，CCF-A 类会议）

[10] Gong T, **Dong Y**, Chen H, Dong B, Li C. Markov subsampling based on Huber criterion[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022 Jul 14;35(2):2250-62.（学生一作，中科院一区 Top）

[11] Lou P, **Dong Y**, Jimeno Yepes A, Li C. A representation model for biological entities by fusing structured axioms with unstructured texts[J]. Bioinformatics, 2021 Apr 15;37(8):1156-63.（中科院三区）

[12] Wu J, **Dong Y**, Gao Z, Gong T, Li C. Dual attention and patient similarity network for drug recommendation[J]. Bioinformatics, 2023 Jan 1;39(1):btad003.（中科院三区）

答辩委员会会议决议

该论文聚焦信息论视角下随机学习算法的泛化理论研究，选题属于学科前沿，具有重要的理论意义和应用前景。该论文取得的主要创新性成果为：

1）基于核化 Rényi 熵的可计算泛化误差估计：提出了新型可计算信息度量准则——核化 Rényi 熵，拓展了现有面向随机迭代学习算法的泛化理论。

2）损失熵诱导的高概率信息论泛化误差上界：提出了新型低维可计算泛化度量——损失熵，构建了更紧致的高概率泛化误差上界。

3）面向多点损失的一致信息论泛化分析框架：提出了信息论视角下的多点学习泛化理论框架，将常见多点学习范式纳入统一的泛化分析框架中。

4）基于信息论的领域泛化理论与算法设计：构建了信息论视角下的领域泛化分析框架，并研发了基于域间分布对齐的领域泛化算法。

论文写作认真，条理清晰，工作量饱满，创新性强，是一篇优秀的博士学位论文。论文工作表明作者已掌握本学科坚实宽广的理论基础和系统深入的专门知识，具备较强的独立从事科研能力。

论文答辩过程中讲述清晰，回答问题正确。经答辩委员会投票表决，一致同意通过董裕欣博士学位论文答辩，并建议授予工学博士学位。

常规评阅人名单

本学位论文共接受 3 位专家评阅，其中常规评阅人 2 名，名单如下：

汤斯亮　教授　　浙江大学

惠维　　教授　西安交通大学

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

（1） 交回学校授予的学位证书；
（2） 学校可在相关媒体上对作者本人的行为进行通报；
（3） 本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。
（4） 本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）：董裕欣　　日期：　2024 年　8 月　16 日

学位论文独创性声明（2）

本人声明：研究生　　所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

（1） 学校可在相关媒体上对本人的失察行为进行通报；
（2） 本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。
（3） 本人接受学校按照有关规定做出的任何处理。

指导教师（签名）：　　日期：　2024 年　8 月　16 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）：董裕欣　　日期：　2024 年　8 月　16 日

指导教师（签名）：　　日期：　2024 年　8 月　16 日