中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

大规模半监督的核方法模型选择研究

作者姓名：李健

指导教师：王伟平 研究员 中国科学院信息工程研究所

学位类别：工学博士

学科专业：网络空间安全

培养单位：中国科学院信息工程研究所

2020 年 6 月

Research on Large Scale Semi-supervised Model Selection for Kernel Methods

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

in Cyberspace Security

By

Li Jian

Supervisor: Professor Wang Weiping

Institute of Information Engineering, Chinese Academy of Sciences

June, 2020

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

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

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

核方法模型选择是核方法理论研究与实际应用的关键，直接决定了核方法的泛化性能。但传统的核方法模型选择方法不适用于当前大规模半监督的数据，存在如下问题：候选核函数集合由人工经验确定，缺乏坚实的泛化理论保证；大多割裂了核方法模型选择、模型训练的过程，求解过程繁琐复杂，缺乏简洁、易求解的模型选择准则；传统核方法模型选择方法内存需求高、计算效率低下，缺乏适用于大规模半监督数据的高效求解算法。

针对核方法模型选择方法处理大规模、半监督数据时存在的不足，本文从理论、准则、算法三个层面展开系统性研究：首先基于谱度量、Rademacher 复杂度、积分算子理论，建立大规模半监督核方法的近似泛化理论；通过最小化理论部分中的泛化误差上界，制定简洁、易求解的核方法模型选择准则；结合常用的大规模加速手段，设计适用于大规模半监督核方法模型选择方法的高效求解算法。本文主要工作和创新点总结如下：

- **大规模半监督的核方法模型选择理论研究。** 通过核矩阵谱分析，为二分类核方法建立基于谱度量的泛化误差界；构造有监督 Rademacher 复杂度、半监督 Rademacher 复杂度的关联，为半监督的多输出学习建立基于局部 Rademacher 复杂度的泛化误差界；基于积分算子理论，使用假设空间容量假设、正则化假设，为结合大规模算法的核岭回归 (KRR) 建立最优泛化误差收敛界。

- **大规模半监督的核方法模型选择准则研究。** 通过最小化基于谱度量的二分类泛化误差界，将最大化谱度量作为模型选择准则；通过最小化基于局部 Rademacher 复杂度的多输出泛化误差界，将最小化核矩阵尾部特征值之和作为模型选择准则；通过深度分析基于 Rademacher 复杂度的多输出泛化误差界，将反向传播更新核超参数作为模型选择准则。

- **大规模半监督的核方法模型选择算法研究。** 本文结合分布式、低秩近似、随机优化等大规模加速手段，设计高效求解算法。使用分布式、随机特征相结合的方法高效求解大规模核岭回归问题；结合 Nyström 采样、预处理共轭梯度下降，设计适用于大规模半监督核岭回归的高效求解算法；使用对偶梯度下降、近端梯度下降等一阶梯度随机优化手段对大规模核方法进行迭代式求解。

总之，针对核方法模型选择面临大规模半监督的瓶颈，本文从泛化理论、模型选择准则、高效算法设计等三个层面递进式展开研究工作。泛化理论结果表明，本文提出的多个近似泛化误差理论均取得了良好的泛化误差收敛率，为核方法模型选择提供了坚实的理论支撑；复杂度分析表明，本文针对核方法模型选择提出的多个求解算法内存需求低、计算高效，提高了核方法模型选择的可用性；实验结果表明，本文提出的多个核方法模型选择方法相比于传统方法，在预测精度或计算效率有显著的提升。本文针对大规模半监督的核方法模型选择展开研究，推动了核方法模型选择泛化理论的发展，提高了核方法在实际应用场景中的可用性，具有重要的理论意义和实用价值。

关键词： 核方法模型选择，半监督核方法，大规模核方法，泛化理论

Abstract

Model selection for kernel methods is the key to the theoretical analysis and practical application of kernel methods, which directly determines the generalization performance of kernel methods. However, conventional model selection methods for kernel methods are not feasible in the current large-scale semi-supervised setting, meanwhile, there are also the following problems: the candidate set of kernels for kernel selection depends on artificial experience, thus current kernel selection methods lack solid generalization theory guarantee; most of traditional kernel selection methods split the process of model selection for kernel methods and model training, leading to a complex selection criterion which is hard to solve, thus kernel selection needs concise and simple criteria; Due to high storage and computational requirements, current methods lack efficient optimization algorithms for larges-scale semi-supervised tasks.

To overcome the bottlenecks of kernel selections in large-scale semi-supervised setting, this dissertation systematically studies three aspects: generalization theory, selection criterion and optimization algorithms. This dissertation first establishes the approximate generalization error bounds for large-scale semi-supervised kernel methods based on the spectral measure, Rademacher complexity and integral operator theory; by minimizing the generalization error bound, this dissertation developed concise and simple kernel selection criteria; combining techniques for solving large-scale problems, this dissertation designed efficient optimization algorithms for large-scale semi-supervised kernel selections. The main work and innovation of this dissertation are as follows:

• **Theoretical research on large-scale semi-supervised model selection for kernel methods.** Using spectral analysis of the kernel matrix, this dissertation derives a spectral measure based generalization error bound for binary classification. Meanwhile, this dissertation bridges the connection between the supervised Rademacher complexity and the semi-supervised Rademacher complexity, proving general local Rademacher complexity based generalization error bounds for semi-supervised vector-valued kernel methods. Moreover, based on the theory of integral operators, this dissertation

makes use of capacity assumption and regularity assumption for hypothesis space, producing the optimal minimax generalization error bounds for large-scale kernel ridge regressions.

• **Criterion research on large-scale semi-supervised model selection for kernel methods.** By minimizing the spectral measure based generalization error bound for binary classification, this dissertation uses the maximization of the spectral measure as a kernel selection criterion. To obtain smaller local Rademacher complexity based multi-class generalization error bounds, this dissertation also employs the minimization of the tail sum of the singular values as a model selection criterion. Through the statistical analysis of Rademacher complexity based vector-valued generalization error bounds, this dissertation devises a criterion that updates kernel hyperparameters by backpropagation w.r.t. the objective.

• **Algorithmic research on large-scale semi-supervised model selection for kernel methods.** In this dissertation, efficient algorithms are designed by combining large-scale acceleration techniques such as distributed learning, low-rank approximation and stochastic optimization. To solve the large-scale kernel ridge regression efficiently, This dissertation combines distributed learning and random features. Combined with Nyström sampling and preconditioned conjugate gradient descent, an efficient algorithm for large-scale semi-supervised kernel ridge regression was presented. This dissertation also uses first-order stochastic optimization methods such as dual gradient descent and proximal gradient descent, to solve kernel methods iteratively.

To sum up, to relieve large-scale semi-supervised bottlenecks in kernel selection, this dissertation focuses on generalization theory, model selection criteria and design of efficient algorithms. Generalization analysis sets up theoretical guarantees for kernel selection, achieving fast convergence rate for generalization error bounds; The analysis of time and space complexity shows the presented algorithms characterize low storage and computational requirements, improving the availability of kernel selection; Extensive experimental results demonstrate that the proposed methods outperform existing kernel selection methods in generalization performance and computational efficiency. The study of the large-scale semi-supervised kernel selection in this dissertation pro-

motes the development of the generalization theory of kernel selection, and improves the usability of kernel methods in practical application scenarios, which has important theoretical significance and practical value.

Keywords: Kernel selection, semi-supervised kernel methods, large-scale kernel methods, generalization theory

目 录

第1章 引言 ····· 1
1.1 研究背景 ····· 1
1.2 研究现状与挑战 ····· 2
1.3 本文工作 ····· 3
1.4 章节安排 ····· 5
第2章 相关工作综述 ····· 7
2.1 核方法 ····· 7
2.2 核方法泛化理论 ····· 9
2.2.1 PAC 学习 ····· 9
2.2.2 VC 维 ····· 10
2.2.3 Rademacher 复杂度 ····· 10
2.2.4 积分算子理论 ····· 12
2.3 核方法模型选择 ····· 13
2.3.1 交叉验证 (CV) ····· 13
2.3.2 近似交叉验证 ····· 14
2.3.3 最大化核对齐值 ····· 14
2.3.4 自动核学习 ····· 14
2.4 半监督核方法 ····· 16
2.4.1 生成式方法 ····· 17
2.4.2 判别式方法 ····· 17
2.4.3 基于图的半监督方法 ····· 18
2.4.4 启发式方法 ····· 19
2.5 大规模核方法 ····· 20
2.5.1 分布式方法 ····· 20
2.5.2 低秩近似方法 ····· 21
2.5.3 随机优化方法 ····· 23
第3章 大规模半监督的核方法模型选择泛化理论研究 ····· 25
3.1 预备知识 ····· 25
3.2 基于谱度量的二分类泛化误差界 ····· 27
3.2.1 谱度量定义 ····· 28
3.2.2 LSSVM 的谱度量泛化误差界 ····· 29

3.2.3 SVM 的谱度量泛化误差界 ····· 31
3.3 基于谱分析的 Rademacher 复杂度泛化理论 ····· 32
3.3.1 Rademacher 复杂度定义 ····· 34
3.3.2 通用的半监督局部 Rademacher 复杂度泛化误差界 ····· 35
3.3.3 半监督核方法的泛化误差界 ····· 38
3.3.4 半监督线性学习器的泛化误差界 ····· 41
3.3.5 相关工作比较 ····· 45
3.3.6 已完成工作对 Rademacher 复杂度泛化理论研究脉络 ····· 48
3.4 基于积分算子理论的最优泛化理论 ····· 49
3.4.1 积分算子理论定义与假设 ····· 50
3.4.2 结合分布式、随机特征的近似核岭回归泛化分析 ····· 51
3.4.3 结合 Nyström 采样、PCG的近似 LapRLS 泛化分析 ····· 57
3.5 本章小结 ····· 60
第4章 大规模半监督的核方法模型选择准则研究 ····· 63
4.1 最大化谱度量 ····· 63
4.1.1 最大化谱度量的特例 ····· 64
4.1.2 与最大均值差异的关联 ····· 65
4.2 最小化核矩阵尾部特征值 ····· 65
4.2.1 多核凸组合方法 (Conv-MKL) ····· 66
4.2.2 多核学习方法 (SMSD-MKL) ····· 68
4.3 反向传播更新核超参数 ····· 69
4.3.1 自动谱核学习 ····· 69
4.3.2 泛化理论保证 ····· 73
第5章 大规模半监督的核方法模型选择算法研究 ····· 75
5.1 常用的大规模核方法加速算法 ····· 75
5.2 结合分治算法、随机特征的核岭回归算法 ····· 77
5.2.1 近似核方法构造 ····· 77
5.2.2 相关工作对比 ····· 78
5.3 结合 Nyström 采样、PCG 加速的半监督核岭回归 ····· 79
5.3.1 近似核方法构造 ····· 79
5.3.2 对比方法介绍 ····· 80
5.4 基于一阶梯度的随机优化算法 ····· 81
5.4.1 对偶梯度下降 ····· 81
5.4.2 近端梯度下降 ····· 84

第6章 实验分析 ····· 89
6.1 大规模半监督核方法模型选择准则 ····· 90
6.1.1 最大化谱度量 ····· 90
6.1.2 最小化核矩阵尾部特征值之和 ····· 92
6.1.3 反向传播更新核超参数 ····· 94
6.2 大规模半监督核方法模型选择算法 ····· 96
6.2.1 结合 Nyström 采样、PCG 加速的半监督核岭回归 ····· 97
6.2.2 结合分治算法、随机特征的核岭回归算法 ····· 100
第7章 总结与展望 ····· 105
参考文献 ····· 107
作者简历及攻读学位期间发表的学术论文与研究成果 ····· 125
致谢 ····· 127

图形列表

1.1 研究内容框架 ····· 4

2.1 半监督支持向量机 (S3VM) 与低密度分割 ····· 18
2.2 分布式机器学习框架 ····· 20

3.1 Rademacher复杂度泛化理论研究脉络 ····· 49
3.2 不同假设对随机特征个数 M、分块数 m、泛化误差收敛率的影响 ···· 52
3.3 无标签数据对分块数、泛化误差收敛率的影响 ····· 54
3.4 大规模半监督核方法泛化理论研究脉络 ····· 60

4.1 自动谱核学习 (ASKL) 框架 ····· 72

6.1 谱度量不同 r 对测试误差的影响 ····· 93
6.2 左图 MNIST 上的准确率曲线、右图 MNIST 上的目标函数曲线 ····· 96
6.3 不同标签比例对测试误差 (RMSE) 的影响 ····· 99
6.4 不同 Nyström 采样比例对测试误差 (RMSE) 的影响 ····· 100
6.5 KRR-DC-RF 在简单回归问题上的表现 ····· 101
6.6 KRR-DC-RF 在困难回归问题上的表现 ····· 102
6.7 KRR-DC-RF 在 covtype 数据集上的测试误差 ····· 103
6.8 KRR-DC-RF 在 SUSY 数据集上的测试误差 ····· 103
6.9 KRR-DC-RF 在 HIGGS 数据集上的测试误差 ····· 104

表格列表

3.1 多输出问题 (VV) 的泛化误差界对比 ····· 46
3.2 多分类问题 (MC) 的泛化误差界对比 ····· 47
3.3 多标签问题 (ML) 的泛化误差界对比 ····· 48

4.1 最大化谱度量与其他核函数选择方法对比 ····· 64
4.2 平稳谱核及其对应谱密度 ····· 70

5.1 近似核岭回归算法对比 ····· 77
5.2 近似核岭回归空间复杂度的对比 ····· 78
5.3 近似核岭回归时间复杂度的对比 ····· 78
5.4 Nyström-PCG 相关算法时间、空间复杂度对比 ····· 81

6.1 最大化谱度量相关方法的平均分类错误率 (%) 对比 ····· 91
6.2 最大化谱度量相关方法的训练时间（秒）对比 ····· 92
6.3 最小化核矩阵尾部特征值之和相关算法的分类准确率 (%) 对比 ····· 94
6.4 自动谱核学习 (ASKL) 对比算法 ····· 95
6.5 自动谱核学习 (ASKL) 相关对比算法的平均测试结果对比 ····· 96
6.6 Nyström-PCG 相关算法平均测试误差对比 ····· 97
6.7 Nyström-PCG 相关算法迭代次数、训练时间对比 ····· 98

主要符号对照表

符号	描述
$\mathbb{R}$、$\mathbb{R}_+$、$\mathbb{R}_+^0$	实数集、正实数集、非负实数集
$\mathbb{N}$、$\mathbb{N}_+$、$\mathbb{N}_+^0$	整数集、正整数集、非负整数集
$\mathcal{X}$、$\mathcal{Y}$、$\mathcal{X}\times\mathcal{Y}$	输入空间、输出空间、样本空间
$\rho_{X\times Y}$、ρ_X	样本空间上联合概率分布、输入空间上边际分布
d、K	输入空间维度、输出空间维度
n、u	有标签样本规模、无标签样本规模
$D^l=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$	有标签数据集
$D^u=\{\boldsymbol{x}_i\}_{i=n+1}^{n+u}$	无标签数据集
m、M、s	分治算法分块数、随机特征维度、Nyström 采样点数
$\kappa:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$	核函数，从 $\mathcal{X}\times\mathcal{X}$ 到实数集 $\mathbb{R}$ 的映射
$\mathcal{H}$	核函数 κ 对应的再生核希尔伯特空间 (RKHS)
$\phi:\mathcal{X}\to\mathcal{H}$	核函数 κ 诱导的隐式特征映射，从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射
$\phi_M:\mathcal{X}\to\mathbb{R}^M$	用于近似核函数 κ 的 M 维随机特征映射
H_κ	使用核函数 κ 对应的假设空间
$f:\mathcal{X}\to\mathcal{Y}$	假设空间中学习器，产生预测标签
$\boldsymbol{W}$、$\|\boldsymbol{W}\|_p$	学习器权重矩阵、假设空间规范化矩阵范数
$\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$	损失函数，用于衡量预测标签与真实标签的差异
$\mathcal{E}(f)$、$\widehat{\mathcal{E}}(f)$	学习器 f 对应的泛化误差、经验误差
$\widehat{f}_n=\arg\min_{f\in H_\kappa}\widehat{\mathcal{E}}(f)$	假设空间中经验误差最小化模型
$f^*=\arg\min_{f\in H_\kappa}\mathcal{E}(f)$	假设空间中泛化误差最小化模型
h^*	真实分布对应的目标假设，存在 $\mathcal{E}(h^*)=0$
$\mathbf{K}$、$\mathbf{N}$	核矩阵、规范化核矩阵 $\mathbf{N}=\mathbf{K}/\lvert\mathbf{K}\rvert_1$
$\mathcal{L}$	假设空间上对应的损失空间
$\widehat{\mathcal{R}}(\mathcal{L})$	损失空间 $\mathcal{L}$ 上经验 Rademacher 复杂度

$\mathcal{R}(\mathcal{L})$	损失空间 $\mathcal{L}$ 上期望 Rademacher 复杂度
$\widehat{\mathcal{R}}(\mathcal{L}_r)$	损失空间 $\mathcal{L}$ 上经验形式的局部 Rademacher 复杂度
$\mathcal{R}(\mathcal{L}_r)$	损失空间 $\mathcal{L}$ 上期望形式的局部 Rademacher 复杂度
$\widehat{\mathcal{R}}(H_r)$	假设空间 H_r 上经验形式的局部 Rademacher 复杂度
$\mathcal{R}(H_r)$	假设空间 H_r 上期望形式的局部 Rademacher 复杂度
λ_j、$\tilde{\lambda}_j$	积分算子对应特征值、权重矩阵 $\boldsymbol{W}$ 对应奇异值
$L^2(\mathcal{X}, \rho_X)$	$\mathcal{X}$ 上的平方可积空间
$L_\kappa : L^2(\mathcal{X}, \rho_X) \to L^2(\mathcal{X}, \rho_X)$	核函数 κ 对应积分算子
λ_A	假设空间正则化 $\|\boldsymbol{W}\|_p$ 参数
λ_I	Laplacian 正则化项参数
$\mathcal{N}(\lambda_A)$	积分算子理论 (Effective Dimension)

第1章　引言

1.1　研究背景

在统计机器学习领域，机器学习旨在基于给定的有限数据样本，通过学习构造出逼近潜在的真实分布（函数依赖关系或内在规律）的学习模型，使得模型具有很强的泛化性能（在未知数据上的预测能力）[1,2]。确定学习模型的过程包括两个步骤：(1) 确定假设空间，(2) 在假设空间中寻找泛化性能最好的模型[3]。由于训练数据是样本空间的有限采样，存在多个假设与数据集一致的情况，因此学习问题是不适定的[1]。机器学习算法在确定学习的过程中对某类假设存在偏好，以唯一确定假设空间，称为“归纳偏置”（inductive bias）[1]。机器学习模型选择就是确定假设空间的过程，而假设空间中假设与问题本身的匹配程度直接决定了学习算法的性能，因而模型选择是机器学习的本质问题。

核方法 (kernel methods) 是统计机器学习中的重要方法，具有坚实泛化理论基础、完备的算法框架，广泛应用于数据挖掘、模式识别、自然语言处理、计算机视觉等各领域[4–6]。核方法研究与应用兴起于 20 世纪末 Cortes 和 Vapnik 提出的支持向量机 (support vector machine, SVM)[7] 及其泛化理论工具：VC 维 (Vapnik-Chervonenkis dimension)[8]。至今核方法仍是是统计机器学习领域研究与应用的热点，如支持向量机、高斯过程 (Gaussian process)[9]、核岭回归 (kernel ridge regression, KRR)[10]、核主成分分析 (kernel principal components analysis, KPCA)[11]、谱聚类 (spectral clustering)[12]等。核方法通过核函数 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{H}}$ 诱导的隐式特征映射 ϕ，将输入样本 $\boldsymbol{x}$ 从输入空间 $\mathcal{X}$ 映射到再生核希尔伯特空间 $\mathcal{H}$ (reproducing kernel Hilbert space, RKHS) 中，获得了强大的特征表示能力；然后在 $\mathcal{H}$ 训练线性学习器，使得线性学习器可以处理非线性问题。核方法对应的假设空间 H_κ 定义为

$$H_\kappa = \left\{ f | f(\boldsymbol{x}) = \langle \boldsymbol{W}, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}}, \quad \forall\, \boldsymbol{x} \in \mathcal{X} \right\},$$

其中隐式特征映射 ϕ 由核函数 κ 唯一确定，因此假设空间与核函数是一一对应的。而核方法模型选择是选择核函数的过程，唯一确定了核方法对应的假设空间 H_κ，因此核方法模型选择直接决定了核方法的学习性能，是核方法理论研究与应用的关键，具有重要的理论意义和实用价值。

1.2 研究现状与挑战

已有核方法模型选择方法主要包括如下三种：

(1) **交叉验证。** 交叉验证方法基于给定的核函数候选集合，每次使用其中的核函数并将样本重复划分为训练集、验证集，使用该核函数重复训练、获得平均验证误差，选取使得验证误差最小的核函数。交叉验证方法在模型选择中得到广泛使用，并发展出多种改进方法，包括：k-折交叉验证[13]、留出法[14]、自助法[15] 等。

(2) **近似交叉验证。** 交叉验证需要在不同划分训练多次，而近似交叉验证只需训练学习器一次，极大的提高了交叉验证进行模型选择的效率。已有多种近似交叉验证方法被提出，包括广义近似交叉验证方法 (GACV)[16]、基于张成界 (span bound) 的近似留一法 (LOO)[17]、k-折交叉验证误差的近似估计[18]、影响函数近似留一法[19,20] 等。

(3) **核对齐方法。** 基于 "相似的输入样本对应标签应该也是相似的" 思想，Cristianini 等提出了最大化核矩阵与标签对齐值的核函数选择方法 (kernel target alignment, KTA)[21]。在此基础上，多种改进方法被提出包括：中心化核对齐 (centered KTA) 准则[22]、基于特征空间的核矩阵评估度量 (feature space-based kernel matrix evaluation, FSM)[23]、核极化 (kernel polarization)[24] 等。

尽管核方法模型选择研究有一定进展，但仍存在很多关键问题亟待解决。概括起来包括以下几点：

(1) **存在大规模、半监督的研究与应用瓶颈**

由于实际应用中无标签数据易于获取，而且数据标注成本较高，训练数据呈现大量有标签、少量无标签的特点。而现有核方法模型选择在监督学习上的理论研究、实际应用，难以扩展到大规模、半监督数据上。

通常核方法的空间复杂度为 $\mathcal{O}(n^2)$、时间复杂度为 $\mathcal{O}(n^3)$，存在大规模瓶颈。而核方法模型选择中需要为每个候选核函数训练模型，令候选核函数集合为 $\mathcal{K}$，则至少重复 $|\mathcal{K}|$ 次模型训练。交叉验证需要重复 k 次重复划分求平均验证误差，因此时间复杂度为 $\mathcal{O}(k|\mathcal{K}|n^3)$；近似交叉验证只需一次划分，因此时间复杂度为 $\mathcal{O}(|\mathcal{K}|n^3)$；而核对齐方法需要计算矩阵-向量乘法，其时间复杂度为 $\mathcal{O}(|\mathcal{K}|n^2)$。同时候选核函数集合 $\mathcal{K}$ 较大，因此现有的核方法模型选择方法计算效率低下，无法为大规模半监督核方法选择核函数。

尽管已有大量半监督核方法的研究，比如直推支持向量机 (transductive support vector machine, TSVM)[25]、基于流形假设的半监督学习[26]、共训练 (co-training)[27] 等方法，但半监督核方法使用传统核方法模型选择方法选取核函数，只考虑了有标签数据而忽略了无标签数据中包含的有用信息。

(2) 缺乏坚实的泛化理论保证

现有核方法模型选择方法需要预先指定候选核函数集合，再从中枚举选出最优核函数。候选核函数集合通常由人工经验给定，无法直接应用现有的核方法泛化理论，因此缺乏坚实的泛化理论保障，模型选择带有盲目性。

(3) 缺乏简洁、易求解的模型选择准则

传统核方法割裂了核方法模型选择、模型训练的过程，导致了制定的核方法模型选择准则存在计算流程繁琐、求解复杂的问题。

(4) 缺乏高效的大规模求解算法

针对核方法面临的大规模瓶颈，已有分布式[28,29]、低秩近似[30,31]、随机优化[32–34] 等多种方法应用到核方法中。但现有核方法模型选择方法没有与上述大规模手段进行结合。而核方法模型选择的空间复杂度、时间复杂度都很高，亟需适用于大规模问题的求解算法。

1.3　本文工作

针对大规模半监督的核方法模型选择存在的问题，本文对大规模半监督的核方法模型选择的理论、准则、算法三个方面递进式展开研究。首先，基于谱度量、Rademacher 复杂度、积分算子等工具，推导出大规模半监督核方法的泛化误差界；其次，通过最小化理论部分的泛化误差界，设计简洁、易求解的核方法模型选择准则；最后，结合分布式、低秩近似、随机优化，设计高效的大规模核方法求解算法。研究内容总体框架如图 1.1 所示。具体内容如下

1. 核方法近似泛化理论

基于核矩阵谱分析得到谱度量，建立谱度量与泛化误差上界的关联，最后推导出适用于二分类支持向量机(support vector machine, SVM)、最小二乘支持向量机 (least square support vector machine, LSSVM) 的泛化误差界。

建立定义在损失空间上 Rademacher 复杂度与假设空间上 Rademacher 复杂度的关联，从而给出半监督核方法的通用泛化误差界；估计半监督核方法的局

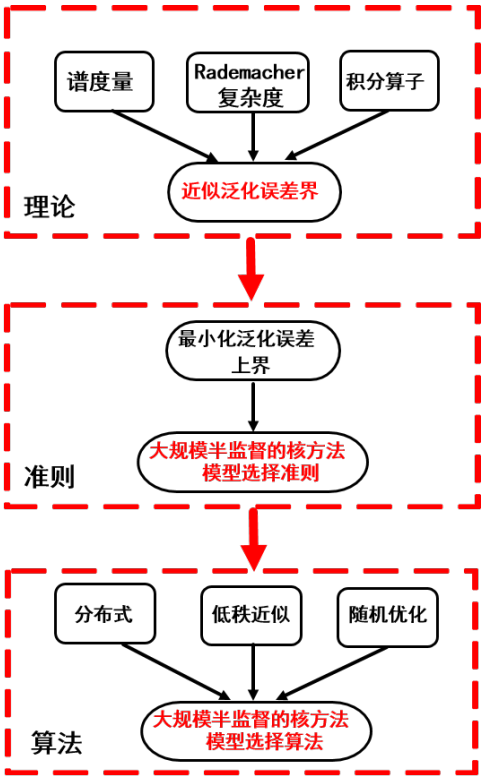

图 1.1 研究内容框架

部 Rademacher 复杂度，给出半监督核方法的泛化误差界；估计半监督线性方法的局部 Rademacher 复杂度，给出半监督线性方法的泛化误差界；与现有的多分类、多标签的核方法和线性方法的泛化误差界进行对比。

基于积分算子理论，使用容量假设、正则化假设，推导出结合分治算法、随机特征的核岭回归的最优泛化误差界；基于积分算子理论，为结合 Nyström 采样、预处理共轭梯度下降算法 (preconditioned conjugate gradient descent, PCG) 的半监督核岭回归方法 (LapRLS) 方法提供泛化理论保证。

2. **模型选择准则**

通过最小化基于谱度量的泛化误差界，将最大化谱度量作为核方法模型选择准则。与其他矩阵分析方法进行关联，与现有核方法模型选择准则对比。

通过最小化基于局部 Rademacher 复杂度的泛化误差界，将最小化核矩阵尾部特征值之和（用于界定局部 Rademacher 复杂度）作为核方法模型选择准则。设计端到端的多核学习方法，同时更新核函数（多核组合系数）、核模型参数。

通过最小化基于 Rademacher 复杂度的泛化误差界，将反向传播更新核超参数作为核方法模型选择准则。使用随机傅里叶特征近似非平稳谱核，构造谱核学习网络，使用反向传播同时更新核函数（谱密度）、核模型参数。

3. 高效求解算法

设计结合分布式、随机特征近似核函数相结合的方法，加速核岭回归(KRR)的求解，分析其在不同假设下对应的分块数、随机特征数、泛化误差收敛率、空间复杂度、时间复杂度。与传统核岭回归方法进行对比，分析分布式、随机特征两种加速手段相结合带来的计算效率提升、泛化性能损失。

设计结合 Nyström 采样近似核矩阵、预处理共轭梯度下降算法 (PCG) 相结合的方法，加速半监督核岭回归 (LapRLS) 的求解。从泛化理论的角度分析该方法所需的 Nyström 采样点个数、PCG 迭代次数，并使用矩阵分块乘法减少内存需求、使用避免矩阵–矩阵相乘的手段提升计算效率。

基于对偶梯度下降方法，设计高效算法求解多核 SVM。基于近端梯度下降算法，设计求解不可微半监督核方法的高效算法。

1.4　章节安排

本文分为七章，具体安排如下：

第一章，引言。介绍大规模核方法的模型选择研究背景与存在问题，在此背景下阐述本文主要内容、创新之处。

第二章，相关工作综述。首先介绍核方法相关概念，再依次介绍核方法泛化理论、核方法模型选择、半监督核方法、大规模核方法等领域的相关工作。

第三章，大规模半监督的核方法模型选择理论研究。分别基于谱度量、局部 Rademacher 复杂度、积分算子等工具，界定核方法泛化误差上界。

第四章，大规模半监督的核方法模型选择准则研究。通过最小化泛化误差上界，制定最大化谱度量、最小化核矩阵尾部特征值之和、反向传播更新核超参数等三个核方法模型选择准则。

第五章，大规模半监督的核方法模型选择算法研究。基于分布式、低秩近似、随机优化等手段，设计求解大规模核方法的高效算法。

第六章，实验分析。通过实验分析，验证大规模半监督核方法模型选择准则、算法的准确性和有效性。

第七章，总结与展望。总结全文工作并展望未来研究工作。

第2章　相关工作综述

2.1　核方法

核方法 (kernel methods) 是一类重要的机器学习方法，具有坚实的理论基础、完备的非线性学习框架，在数据挖掘和模式识别等领域得到广泛应用[4–6]。核方法研究与应用的基础是再生核理论，而再生核理论早在上世纪50年代就已被提出、并得到广泛研究[35]。再生核理论中 Moore-Aronszajn 定理[35] 指出正定的核函数 κ 对应的再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS) $\mathcal{H}$ 存在且唯一的。核方法通过特征映射 $\phi : \mathcal{X} \to \mathcal{H}$ 将输入样本从输入空间 $\mathcal{X}$ 映射到 $\mathcal{H}$ 中，然而由于 RKHS (高维甚至无穷维的空间) 中计算内积等操作过于复杂，最开始核方法没有得到广泛应用。

直到 20 世纪末，Cortes 和 Vapnik 在 SVM[7] 中使用核技巧 (kernel trick) 通过映射的内积形式将输入样本隐式地映射到RKHS 中，无需显式地给出特征映射 ϕ 的具体形式、避免了 RKHS 中的运算。其中，Mercer 定理[36] 指出：只要核函数 $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 是对称的正定函数，一定存在核函数 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{H}}$ 诱导的隐式特征映射 ϕ、对应的再生核希尔伯特空间 $\mathcal{H}$。通过使用核技巧，只需构造对称的正定核函数 κ，即可避免使用显式特征映射、高维空间的内积运算，并能够使用线性算法解决非线性问题，极大地提高了核方法的普适性。由此，核方法成为统计机器学习研究与应用的热点，掀起了将已有线性方法核化的研究热潮，如高斯过程 (Gaussian process)[9]、核岭回归 (kernel ridge regression, KRR)[10]、核主成分分析 (kernel principal components analysis, KPCA)、[11]、核判别分析 (kernel Fisher discriminant analysis, KFDA)[37]、谱聚类 (spectral clustering)[12]等，显著增强了以线性投影为基础的算法处理非线性问题的能力。核方法在大量应用领域已取得了引人注目的成就，如文本分类、语音识别、基因表达等。

下面介绍核方法中使用到的核函数、积分算子、假设空间、泛化误差等关键概念。令 $\mathcal{X}$ 为输入空间、 $\mathcal{Y}$ 为输出空间、$\mathcal{X} \times \mathcal{Y}$ 为样本空间。以监督学习中的二分类问题为例，其对应输出空间为 $\mathcal{Y} = \{+1, -1\}$。样本集合 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ 独立同分布 (identically independently distributed, i.i.d.) 地采样于样本空间 $\mathcal{X} \times \mathcal{Y}$ 上一个固定但未知的概率分布 $\rho_{X \times Y}$。核函数 $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 为映射的内积形式，

对称的正定核函数

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{H}},$$

隐式地诱导了特征映射 $\phi : \mathcal{X} \to \mathcal{H}$。训练数据集上的 $n \times n$ 的核矩阵 $\mathbf{K} = \left[\frac{1}{n}\kappa(\boldsymbol{x}, \boldsymbol{x}')\right]_{i,j=1}^{n}$ 为半正定核矩阵[4,35]。将积分算子定义为

$$(L_\kappa g)(\boldsymbol{x}) = \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{z}) d\rho_X(\boldsymbol{z}),$$

其中 ρ_X 是在输入数据 $\mathcal{X}$ 上的边际分布。核矩阵 $\mathbf{K}$ 是积分算子 L_κ 在样本集合 D 上的经验形式[38,39]。 Mercer 定理[35,40] 说明核函数 κ 可以由它对应积分算子的特征值 λ_j、特征向量 φ_j 线性表征

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{\infty} \lambda_j \varphi_j(\boldsymbol{x}) \varphi_j(\boldsymbol{x}').$$

核方法通过由核矩阵 κ 诱导的隐式特征映射将输入映射到 RKSH $\mathcal{H}$ 中，再在 $\mathcal{H}$ 使用线性学习器进行学习。假设空间定义为

$$H_\kappa = \left\{ f \mid \boldsymbol{x} \to f(\boldsymbol{x}) = \langle \boldsymbol{W}, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}} : \|\boldsymbol{W}\|_p \leq 1 \right\},$$

其中，$\boldsymbol{W} \in \mathcal{H} \times \mathcal{Y}$ 为待学习的模型参数。机器学习模型学习目标为获得更好的泛化性能（即在未知数据上的预测性能），模型泛化性能通常由泛化误差 (generalization error) 进行衡量

$$\mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(\boldsymbol{x}), y) \, d \, \rho_{X \times Y}(\boldsymbol{x}, y),$$

其中，$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 为损失函数，衡量了预测输出与实际输出的差异。因此理想的机器学习方法是，以最小化泛化误差作为优化目标进行求解模型。但由于概率分布 $\rho_{X \times Y}$ 未知，无法对模型的泛化误差 $\mathcal{E}(f)$ 进行评估。只能使用样本集 D 上经验误差 (empirical error) 进行估计，并使用经验误差最小化 (empirical risk minimization, ERM) 作为训练目标

$$\arg\min_{f \in H_\kappa} \widehat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i).$$

为避免过拟合，实际模型求解中通常加入正则化项，从而权衡模型方差 (variance)、模型偏差 (bias)[41]。常见正则化核学习算法包括支持向量机 (support

vector machine，SVM)[1]、核岭回归 (kernel ridge regression，KRR)[42]和最小二乘支持向量机 (least square support vector machine，LSSVM)[43]等。核方法通过核函数 κ 诱导的特征映射 ϕ 将输入空间 $\mathcal{X}$ 中数据映射到高维（甚至无穷维）的 RKHS $\mathcal{H}$，然后在 $\mathcal{H}$ 中学习线性学习器，使相对简单的线性方法具有了非常强的表达能力[44]。由于核方法的良好的泛化性能，已经在很多领域得到广泛应用。为方便核方法的研究与应用，研究者们开发了多个核方法工具包，包括 LIBSVM[45]、LIBLINEAR[46]、SVM Light [1]、Shogun Toolbox [2] 等。

2.2　核方法泛化理论

机器学习模型的泛化理论分析用于获取假设空间中泛化误差 $\mathcal{E}(f)$ 最小的模型[14]。然而由于概率分布未知，无法直接计算模型的泛化误差，通常利用经验误差并结合模型复杂度建立泛化误差上界，通过最小化泛化误差上界进行泛化误差分析[47,48]。界定泛化误差上界的关键是如何度量假设空间复杂度，不同的模型复杂度对应不同的泛化误差界。常用的假设空间复杂度度量工具包括： VC 维 (VC dimension)[3]、半径间隔界 (radius-margin)[3]、Rademacher 复杂度[49]、算法稳定性 (stability)[50,51]、覆盖数 (covering number)[52]、压缩系数 (compression coefficient)[53] 等。本节主要从 PAC 学习出发，对 VC 维、Rademacher 复杂度、积分算子理论进行介绍。

2.2.1　PAC 学习

计算学习理论的基础是概率近似正确 (probably approximately correct, PAC) 学习理论[54]。真实分布对应假设为 h^*，是由概率分布 $\rho_{X\times Y}$ 决定的输入空间 $\mathcal{X}$ 到输出空间 $\mathcal{Y}$ 的映射，并有 $\mathcal{E}(h^*)=0$。 PAC 辨识 (PAC Identify) 定义为：对于 $\epsilon,\delta\in(0,1)$，使用学习算法 $\mathcal{A}$ 学得模型 $f\in H_\kappa$ 满足

$$\Pr(\mathcal{E}(f)\le\epsilon)\ge 1-\delta,$$

则称学习算法 $\mathcal{A}$ 能从假设空间 H_κ 辨识最优假设 h^*。也就是说学习算法 $\mathcal{A}$ 能以较大概率 (至少 $1-\delta$) 学得最优假设 h^* 的近似（误差最多为 ϵ）。 PAC 学习给出了刻画学习算法泛化能力的框架，由于假设空间 H_κ 中包含了学习算法所有可能的输出，因此度量假设空间复杂度是 PAC 学习的关键。

[1] http://svmlight.joachims.org/

[2] https://www.shogun-toolbox.org/

对于假设空间有限的学习任务，PAC 学习能够直接分析学习算法的学习能力；但对于学习任务面临无限假设空间的时候，需要额外引入工具度量假设空间复杂度，进而分析学习器的泛化能力。

2.2.2 VC 维

VC 维 (Vapnik-Chervonenkis dimension) 最初由 Vapnik 和 Chervonenkis 给出定义[8]，是分布无关 (distribution-free)、数据独立 (data-independent) 的假设空间复杂度度量工具，常用于分析核方法的泛化能力。

对于 $n \in \mathbb{N}_+$，假设空间 H_κ 的增长函数 $\Pi_{H_\kappa}(n)$ 定义为

$$\Pi_{H_\kappa}(n) = \max_{\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}\in\mathcal{X}} \left|\left\{\left(f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_n)\right)|f \in H_\kappa\right\}\right|.$$

增长函数 $\Pi_{H_\kappa}(n)$ 表示假设空间 H_κ 对于 n 个样例能赋予标签的可能数。假设空间的 VC 维定为能被 H_κ 打散的最大样例集的规模

$$\mathrm{VC}(H_\kappa) = \max\{n : \Pi_{H_\kappa}(n) = 2^n\}.$$

VC 维常用于衡量学习器期望误差与经验误差的差异：对于任意 $f \in H_\kappa$, $\delta \in (0,1)$，以至少 $1-\delta$ 的概率存在[3]

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \le \sqrt{\frac{1}{n}\left[D\left(\log\left(\frac{2n}{D}\right)+1\right) - \log\left(\frac{\delta}{4}\right)\right]},$$

其中 $D = \mathrm{VC}(H_\kappa)$ 为 VC 维，$\mathcal{E}(f)$ 为泛化误差，$\widehat{\mathcal{E}}(f)$ 为经验误差。

2.2.3 Rademacher 复杂度

VC 维的泛化误差界是分布无关的 (distribution free)，因此具有普适性；同时 VC 维是数据独立的 (data-independent)，没有考虑数据自身。分布无关、数据独立的特性使得基于 VC 维得到的泛化误差界通常过于保守。

对统计学习理论深入研究后，Shawe-Taylor 注意到数据依赖 (data-dependent) 假设空间复杂性度量的重要性[55]。由于发现 Rademacher 复杂度在对假设空间进行复杂性度量时，依赖特定的样本集（数据分布），Koltchinskii 等将 Rademacher 复杂度引入到统计学习理论中，用于分析学习模型泛化能力[49,56]。 Rademacher 复杂度是分布相关、数据依赖的，能够以比 VC 维更紧致的方式关联经验过程，从而对学习模型的泛化性能分析有较大提升、泛化误差收敛率也稍快。

对于给定样本集合 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \in \mathcal{X} \times \mathcal{Y}$, 假设空间 H_κ 的经验Rademacher复杂度定义为

$$\widehat{\mathcal{R}}(H_\kappa) = \frac{1}{n}\,\mathbb{E}_\epsilon\left[\sup_{f\in H_\kappa}\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right],$$

其中，$\epsilon_1, \cdots, \epsilon_n$ 是独立地采样于 Rademacher 分布的随机变量，满足 $\Pr(\epsilon_i = +1) = \Pr(\epsilon_i = -1) = 1/2, \forall\, i = 1, 2, \cdots, n$。期望 Rademacher 复杂度与概率分布 $\rho_{X\times Y}$、采样样本相关，定义为

$$\mathcal{R}(H_\kappa) = \mathbb{E}_{D\sim\rho_{X\times Y}}[\widehat{\mathcal{R}}(H_\kappa)].$$

Rademacher 复杂度可以用来衡量采样样本 D 的表达能力（经验误差与泛化误差的差异上界）[57]：

$$\mathbb{E}_{D\sim\rho_{X\times Y}}\left[\sup_{f\in H_\kappa}\left(\mathcal{E}(f) - \widehat{\mathcal{E}}(f)\right)\right] \le 2\mathcal{R}(H_\kappa),$$

其中采样样本的表达能力 $\sup_{f\in H_\kappa}\left(\mathcal{E}(f) - \widehat{\mathcal{E}}(f)\right)$ 越小越好，代表了真实误差与测试误差相差不多，相当于避免了过拟合。

以二分类问题 $\mathcal{Y} = \{+1, -1\}$ 为例，存在如下数据依赖的泛化误差界[49,58]：对于任意 $\delta \in (0, 1), f \in H_\kappa$，以至少 $1 - \delta$ 的概率存在

$$\mathcal{E}(f) - \widehat{\mathcal{E}}(f) \le 2\widehat{\mathcal{R}}(H_\kappa) + 4\sqrt{\frac{2\log(4/\delta)}{n}}.$$

上述不等式中需要估计经验 Rademacher 复杂度 $\widehat{\mathcal{R}}(H_\kappa)$，通常得到收敛率为 $\mathcal{O}(1/\sqrt{n})$ 的上界。 Massart 引理[59] 建立了 VC 维与 Rademacher 复杂度的联系

$$\widehat{\mathcal{R}}(H_\kappa) \le \sqrt{\frac{2D\log en/D}{n}},$$

其中 $D = \mathrm{VC}(H_\kappa)$ 为假设空间 $\mathcal{H}$ 对应的 VC 维。文献[60] 给出了 Rademacher 复杂度与覆盖数之间的关系，Srebo 等在文献[61,62] 中进行了进一步改进。

Bartlett 等在 Rademacher 复杂度的研究中发现，对学习模型泛化能力起关键作用不是整个假设空间中的函数, 而是假设空间中具有较小方差的函数所构成的子空间[63–66]，从而提出了局部 Rademacher 复杂度 (local Rademacher complexity) 的概念。在某些条件下（如核矩阵特征值呈指数级下降）[47,65]，使用局部 Rademacher 复杂度获得的泛化误差收敛率能够达到 $\mathcal{O}(1/n)$，而全局 Rademacher 复杂度对应的泛化误差收敛率始终为 $\mathcal{O}(1/\sqrt{n})$。局部 Rademacher 复杂度是本文泛化理论研究、设计核方法模型选择准则的核心工具。

2.2.4 积分算子理论

积分算子 (integral operator) 理论是研究核方法泛化理论的重要工具[39,67–70]。Zhang 最先考虑利用积分算子代替覆盖数研究核方法学习理论[67]。Smale 和 Zhou 用积分算子界定了经验数据上学得假设与最优假设之间偏差上界[68,69,71]。De Vito 等应用积分算子理论给出 KRR 算法的泛化误差上界，并将正则化参数选择问题归约为求解偏差-方差分解问题 (bias-variance)[72]。

积分算子理论通过定义在积分算子上的有限维 (effective dimension) 衡量假设空间复杂度。给定积分算子 L_κ，有效维定义为

$$\mathcal{N}(\lambda_A) = \mathrm{Tr}\left((L_\kappa + \lambda_A I)^{-1} L_\kappa\right), \quad \lambda_A > 0.$$

Caponnetto 和 Vito 为积分算子理论引入两个重要假设[70]：容量假设（capacity assumption, 假设 3.14）用于限制假设空间大小，正则化假设（regularity assumption，假设 3.15）用于对假设空间进行正则化。基于容量假设、正则化假设，可以获得为核岭回归 (KRR) 问题最优泛化误差收敛率[69,70]

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

其中 $\gamma \in [0, 1]$ 控制着假设空间 H_κ 的大小，$r \in [1/2, 1]$ 控制假设空间H_κ 的正则化程度。 γ、r 的取值是由学习任务本身的难度潜在决定，而积分算子理论总能获得对应核岭回归的最优泛化误差收敛率。该最优泛化误差收敛率根据 γ、r 的不同取值，在 $\mathcal{O}(1/n)$ 与 $\mathcal{O}(1/\sqrt{n})$ 之间。

之后，大量工作基于积分算子理论对核岭回归近似算法的最优泛化理论展开研究： Rudi 等基于积分算子理论，使用容量假设、正则化假设为 Nyström 近似 KRR 推导出最优泛化收敛率[73]，之后也给出使用随机特征近似假设 KRR 的最优泛化误差界[74]。 Zhang 等基于积分算子理论，给出了分治核岭回归最优泛化误差收敛率[28]，但其分析没有使用经典的容量假设、正则化假设；而 Lin 等基于经典的积分算子最优率框架（使用容量假设、正则化假设），获得了分治核岭回归的最优误差收敛率[75,76]。基于积分算子理论，使用随机梯度方法求解核岭回归的最优泛化误差收敛率也得到广泛研究[29,77]。近期，多种加速手段相结合的 KRR 对应的最优泛化理论成为研究热点[78–82]。

2.3　核方法模型选择

在（半）监督学习框架中，训练样本是由某个函数 h^* 依赖关系产生的输入输出对。而机器学习的目的是找到某个函数 f，使其尽量逼近目标函数 h^*，该过程可以概括为两步：

- **模型选择**：确定要学习的函数集（通常称为假设空间，$\mathcal{H}$）。
- **模型训练**：从假设空间 $\mathcal{H}$ 中找出泛化误差最小的模型 f^*。

但由于产生训练样本的潜在概率分布未知，无法直接计算泛化误差，实际模型训练中基于训练数据找出经验误差最小的模型 $\widehat{f}_n$。统计学习理论通常衡量经验误差最小化模型与真实模型之间泛化误差的差异，存在如下误差分解[2,83]：

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(h^*) = \underbrace{\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*)}_{\text{估计误差}} + \underbrace{\mathcal{E}(f^*) - \mathcal{E}(h^*)}_{\text{逼近误差}}.$$

模型选择的过程是减少逼近误差的过程，而模型训练是减小估计误差的过程。核学习算法的假设空间 H_κ 是由核函数 κ 唯一确定[35]，因此核函数直接决定核学习算法的性能[84]。下面介绍核方法学习的两个步骤

- **核方法模型选择**：为核方法选择对应核函数，从而确定假设空间 H_κ，是降低逼近误差的过程；
- **核方法模型训练**：通过经验误差最小化 (ERM) 训练模型参数，从而在假设空间 H_κ 中选取经验最小化学习器 $\arg\min_{f\in H_\kappa} \widehat{\mathcal{E}}(f)$，是降低估计误差的过程。

可以看出核学习模型的泛化性能，很大程度上取决于其对应假设空间 H_κ，因此核方法模型选择至关重要，但遗憾的是核函数的选择是一个未决的问题[85]。

2.3.1　交叉验证 (CV)

核方法模型选择中最常用的方法是交叉验证 (cross-validation, CV)，包括：k-折交叉验证[13]、留出法[14]、自助法[15] 等多种方法。这些方法都是将训练数据划分为训练集、验证集，为每个候选核函数在训练集上训练核学习器，在验证集上评估该候选核学习器对应的验证误差；重复上述过程多次，最终选取使得平均验证误差最小的核函数。虽然这些方法得到广泛应用，但存在如下缺点

(1) **缺乏坚实的理论基础**。自助法、k-折交叉验证不是泛化误差的无偏估计，因此在理论上无法保证泛化性能[86]。留出法虽然是泛化误差的无偏估计[14]，

但方差较大，容易出现过拟合问题。

(2) **计算复杂度较高**。选取核函数需要重复训练核学习器（假设为 k 次），而训练核学习器的时间复杂度通常为 $\mathcal{O}(n^3)$，因此训练候选核函数集中单个核函数时间复杂度为 $\mathcal{O}(kn^3)$。

2.3.2 近似交叉验证

为提高交叉验证效率，多种近似交叉验证方法被提出[14,19,20,87,88]。Wahba 为 SVM 提出了广义交叉验证方法 (GACV)[16]。Vapnik 和 Chapelle 提出了基于张成界 (span bound) 的近似留一法 (LOO) 的误差估计[17]。Cawley 等将上述留一法近似方法推广到稀疏最小二乘支持向量机 (sparse LSSVM)[88]。An 等给出了在 LSSVM 和 SVM 上的 k-折交叉验证误差的近似估计方法[18]。Debruyne 等基于影响函数 (influence function) 概念，并采用泰勒展开近似留一法 (LOO)[19,20]。对于一个候选核函数，近似交叉验证方法只需训练一次核学习器，存在计算效率较低 $\mathcal{O}(n^3)$、缺乏坚实理论保证、结果不够稳定可靠等问题。

2.3.3 最大化核对齐值

核对齐 (kernel target alignment, KTA) 方法最早由 Cristianini 和 Shawe-Taylor 提出[89]。该方法无需训练核学习器，而是通过最大化核矩阵与标签的对齐值来选取最优核函数

$$\arg\max_{\kappa\in\mathcal{K}} \ \mathbf{y}^{\top}\mathbf{K}\mathbf{y},$$

其中 $\mathbf{K}$ 为核函数 κ 对应的核矩阵，$\mathbf{y}=(y_1,\cdots,y_n)^{\top}$ 为标签向量。为改进核对齐性能，Cortes 等提出了中心化核对齐 (centered KTA) 准则[22]：

$$\arg\max_{\kappa\in\mathcal{K}} \ \mathbf{y}^{\top}\mathbf{K}_c\mathbf{y},$$

其中 $\mathbf{K}_c$ 为中心化核矩阵。在原始核对齐方法 (KTA) 的基础上，基于特征空间的核矩阵评估度量 (feature space-based kernel matrix evaluation, FSM)[23]、核极化 (kernel polarization)[24] 进一步改进核矩阵、标签值对齐的方法。

2.3.4 自动核学习

交叉验证、近似交叉验证、核对齐等传统核方法模型选择方法将核方法模型选择（确定假设空间）、训练核学习器（在假设空间中选取最优假设）分隔

开，分成两步进行。首先在候选核函数中选取最优泛化性能的核函数；然后基于该核函数，学习训练数据集 D 上的经验损失最小化的模型。

这两种方法存在以下问题：

(1) 需要预先给出候选核函数集合，而候选核函数范围依赖于人工经验。

(2) 对于每个候选核函数，都需要训练核学习器至少一次。当候选核函数集合较大时，计算效率低下。

(3) 无法直接应用最小化泛化误差界理论，缺乏泛化理论支持。

而自动核学习 (kernel learning) 方法将核方法模型选择、训练核学习器两个步骤合并，以最小化经验误差为优化目标，通过端到端的形式同时学习核函数、核模型参数。自动核学习存在如下好处：

(1) 无需预先给出候选核函数集合。只需给出初始核超参数，训练过程中自动更新核超参数，以获得更优的核超参数。

(2) 只需训练核学习器一次。在训练过程中针对优化目标，以端到端形式同时优化核函数、核学习器参数。

(3) 可以直接应用泛化误差理论，使用泛化理论指导核学习算法的设计。

2.3.4.1　多核学习 (MKL)

多核学习 (multiple kernel learning, MKL) 同时学习多个核函数的凸组合系数、核学习器模型[90]，其核函数是多个基核的凸组合

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{m} \mu_m \kappa_m(\boldsymbol{x}, \boldsymbol{x}'),\ \mu_i \geq 0,\ \sum_{i=1}^{m} \mu_i = 1,$$

其中，$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)$ 为多核组合系数。多核函数设计更合理的优化方程提高泛化能力，包括 ℓ_p 范数约束[91]、非线性多核组合[92,93]、融入半径信息的核学习方法[48,94]、融入局部 Rademacher 复杂度信息[47] 等。为求解复杂的多核学习问题，多种高效优化算法已被提出，包括半无限规划 (semi-infinite linear programming, SILP)[95]、随机梯度下降方法[96,97] 等。

Argyriou 和 Micchelli 提出了基于 DC 算法的无限核学习方法[98]。 Gehler 和 Nowozin 设计了一种求解无限核学习 (infinite kernel learning) 方法的框架[99]，并利用半无限规划求解该问题。

2.3.4.2 自动谱核学习

基于 Bochner 定理， Rahimi 和 Recht 使用蒙特卡洛近似得出的随机傅里叶特征近似平移不变核[31]。 Zhang 等提出使用随机傅里叶特征构造平移不变核网络，并通过反向传播更新谱密度，从而同时学习谱密度、模型参数[100]。

而 Yaglom 定理说明，能够使用组合形式的随机傅里叶随机特征近似任意谱核函数[101]。 Yaglom 定理为：当且仅当一般核函数 $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ 满足以下形式

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{i(\omega^\top \boldsymbol{x} - \omega'^\top \boldsymbol{x}')} \mu(d\omega, d\omega'),$$

该核函数在输入空间 $\mathcal{X}$ 是正定的。其中 $\mu(d\omega, d\omega')$ 为与半正定 (positive semi definite, PSD) 谱度量 $s(\omega, \omega')$ 相关的 Lebesgue-Stieltjes 度量。

Remes 等使用随机傅里叶随机特征构造有限维的谱核网络，并使用反向传播更新核函数对应的谱密度，以端到端的形式同时学习核函数、核学习器[102]。谱核网络的使用，只需更新单个核函数，避免了使用自动核学习中多核组合的形式。谱核学习目前广泛应用于高斯过程的学习[102–104]、贝叶斯推断[105]。

Xue 等将谱核网络引入到一般的核方法中，并使用深度网络[106]。已完成工作使用 Rademacher 复杂度为谱核网络提供了泛化理论保证，并指导学习框架的设计[107]，之后将可解释的谱核网络扩展到多层卷积网络[108]。

2.4 半监督核方法

真实世界学习任务中由于无标签数据容易获取，而数据标注成本较高，通常情况下训练数据由大量无标签、少量有标签数据组成。若使用传统监督学习方法，仅能利用少量有标签数据进行训练，摒弃了大量无标签数据中包含的有用信息，通常模型泛化性能较差。

半监督学习 (semi-supervised learning, SSL) 基于少量有标签样本及大量无标签样本的训练集进行学习[109]，半监督学习的关键是利用无标签样本提升学习器性能。对于半监督问题，有标签训练集合 $D^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ 中样本独立同分布 (i.i.d.) 地采样于分布 $\rho_{X\times Y}$，无标签数据集合 $D^u = \{\boldsymbol{x}_i\}_{i=1}^u \in \mathcal{X}$ 中输入样本独立同分布地采样于边际概率分布 ρ_X。在常见的半监督应用中，有标签样本数远远少于无标签样本数，即 $n \ll u$。半监督学习方法包括：生成式方法 (generative models)[110–112]、半监督 SVM (semi-supervised support vector ma-

chine, S3VM)[25,113,114]、基于图的半监督学习 (graph-based semi-supervised learning, GSSL)[26,115]、启发式方法(heuristic approaches)[27,116] 等四类。

2.4.1　生成式方法

生成式方法 (generative models) 是直接基于生成式的模型，生成式方法假定所有数据都是由同一个潜在的概率分布 $\rho_{X\times Y}$ 生成的。基于上述假设可以通过潜在的模型参数将无标签数据与学习目标关联起来，并将缺失的标签视为隐变量，使用 EM 算法进行求解。

生成式方法预测函数为：$\arg\max_y\ p(y|\boldsymbol{x},\Theta)$，其中 Θ 为概率分布参数。基于有标签数据集、无标签数据集，对概率分布参数 Θ 进行学习

$$\arg\max_{\Theta}\ \Big(\log p\big(\{\boldsymbol{x}_i,y_i\}_{i=1}^{n}|\Theta\big)+\lambda_I\log p\big(\{\boldsymbol{x}_i\}_{i=n+1}^{n+u}|\Theta\big)\Big),$$

其中，λ_I 用于权衡有标签数据、无标签数据的影响。此类方法依赖于潜在分布的假设，不同假设对应于不同方法。已有半监督生成式方法将潜在数据分布视为高斯模型[117]、高斯混合模型 (Gaussian mixture model, GMM)[110,111]、混合专家模型 (mixture of experts)[112]、朴素贝叶斯模型 (naïve Bayes models)[111]等。

如果假设分布与实际分布相符，使用无标签数据比只使用有标签数据能获得更好的泛化性能[118]；而如果假设分布于实际分布不符，使用无标签数据会降低泛化性能，半监督学习器精度低于只使用有标签数据的监督方法[119]。

2.4.2　判别式方法

判别式方法通过最大间隔法同时训练有标签样本、无标签样本的学习决策边界。基于低密度分隔 (low-density separation) 假设，决策边界穿过低密度区域的同时，使得不同标签样本间隔最大[120,121]。如图 2.1 所示（源自[122]），半监督支持向量机 (S3VM) 对支持向量机 (SVM) 的划分超平面进行了修正，使其穿过低密度区域。判别式方法应用到半监督学习的方法包括费舍尔线性判别法 (Fisher Linear Discriminative Analysis, FDA)[123]、半监督支持向量机 (Semi-supervised Support Vector Machine, S3VM)[25]、熵正则化[124]等。

其中的典型方法是半监督 SVM 中，由 Joachims 提出的直推式支持向量机 (Transductive Support Vector Machine, TSVM)，其改进方法包括 S4VM[125]、

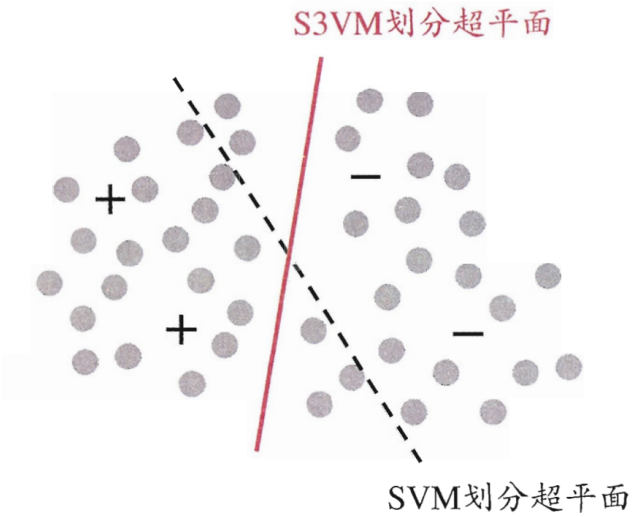

图 2.1 半监督支持向量机 (S3VM) 与低密度分割

PTSVM[126]。以半监督二分类为例，TSVM 的学习目标为

$$\arg\min_{f\in H_\kappa}\left(\sum_{i=1}^{n}(1-y_i f(\boldsymbol{x}_i))_+ + \lambda_A\|f\|_{\mathcal{H}}^2 + \lambda_I\sum_{i=n+1}^{n+u}(1-|f(\boldsymbol{x}_i)|)_+\right),$$

其中，λ_I 用于权衡有标签损失、无标签正则化项的作用。由于 $(1-|f(\boldsymbol{x}_i)|)_+$ 的存在，上式的求解是非凸的，难以找到全局最优解。因此半监督 SVM 研究的一个重点是如何设计出高效的优化求解策略，由此发展处基于图核 (graph kernel) 函数梯度下降的 LDS[113]、基于标记均值估计的 meanS3VM[114] 等方法。

2.4.3 基于图的半监督方法

基于图的半监督学习方法 (graph-based semi-supervised learning, GSSL) 假设数据满足流形假设 (manifold assumption)：输入数据能够近似地嵌入到低维流形空间中。基于流形假设，对所有数据构造图表达 $\boldsymbol{S}$，图中顶点表示输入样本，图中边表示两个样本之间的相似度。图表达 $\boldsymbol{S}$ 实际上是所有输入样本上的相似度矩阵，构造 $\boldsymbol{S}$ 首先使用 k-近邻算法为每个样本找到最相似的 k 个点；再使用相似度度量（如热核函数）求出与之最接近 k 个样本的相似度，将该相似度作为矩阵元素值；将矩阵中相似度较小两个样本对应元素值设置为 0。

GSSL 的本质是通过最小化能量函数 (energy function) 实现标签传播 (label propagation)[115,127]。能量函数的定义为

$$\sum_{i,j=1}^{n+u}\boldsymbol{S}_{ij}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2 = \mathbf{f}^T\mathbf{L}\mathbf{f},$$

其中 $\mathbf{f} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \cdots, f(\boldsymbol{x}_{n+u}))^\top$。 $\mathbf{L}$ 为 Laplacian 矩阵定义为 $\mathbf{L} = \mathbf{D} - \boldsymbol{S}$，其中对角矩阵的对角元素为 $\mathbf{D}_{ii} = \sum_{j=1}^{n+u} \boldsymbol{S}_{ij}$。从上式可以看出最小化能量函数，能够使得相近样本的标签也越相似，从而达到了标签传递的目的。Blum 和 Chawla 最早提出基于图的半监督方法 – 最小割 (mincut)，并给出了能量函数的定义[127]。之后，Zhu 等提出比例割法 (ratio cut)、Zhou 等提出归一化割法 (normalized cut)、Zhu 等提出调和函数法 (harmonic cut) 对最小割方法存在的问题进行改进。

基于 Laplacian 算子等工作[128,129]，Belkin 等归纳出流形正则化 (manifold regularization) 学习框架[26]，将能量函数放在优化目标中，与经验损失、正则化项共同进行学习。流形正则化 (manifold regularization) 学习框架为

$$\arg\min_{f \in H_\kappa} \left(\frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i) + \lambda_A \|f\|_{\mathcal{H}}^2 + \lambda_I \mathbf{f}^T \mathbf{L} \mathbf{f} \right).$$

与监督核方法类似，流形正则化也存在大规模瓶颈。针对流形正则化的大规模问题，Fergus 等提出 Eigenfunction 方法[130]，使用 Laplacian 矩阵中少量的特征向量表示潜在的流形结构； Zhang 等提出最小树割方法 (minimum tree cut, MTC)[131]，使用生成树近似 Laplacian 矩阵，并用最小化树割打标签； Liu 等提出锚点正则化 (anchor graph regularization, AGR)[132,133], 选取无标签的锚点数据，能够保证 Laplacian 矩阵是半正定的，然后使用锚点、有标签数据进行训练。Wang 等人提出使用局部锚点嵌入[134]，加速了锚点图的重构过程。在 AGR 工作中需要首先使用 k-means 寻找聚类中心点作为锚点，而已完成工作使用 Nyström 采样直接将均匀采样的样本点作为锚点[135]，无需进行效率低下的 k-means 步骤，同时能够达到相似的泛化误差理论保证、学习性能。

2.4.4　启发式方法

启发式方法 (heuristic approaches) 主要包括自训练 (self-trianing)[136]、基于差异的方法 (disagreement-based methods)[27]。

自训练方法 (self-trianing) 是由 Rosenberg 等提出[136]：先使用有标记数据训练出模型，对未标记数据进行预测，并将预测结果中最有把握的结果作为标记，并重复此训练过程。

基于分歧的方法(disagreement-based methods) 使用多个学习器，而通过学习器之间的分歧 (disagreement) 达到对利用未标记数据的目的。比如共训练 (co-training)[27,116]，针对多视角 (multi-view) 训练多个学习器，将最有把握的标记提

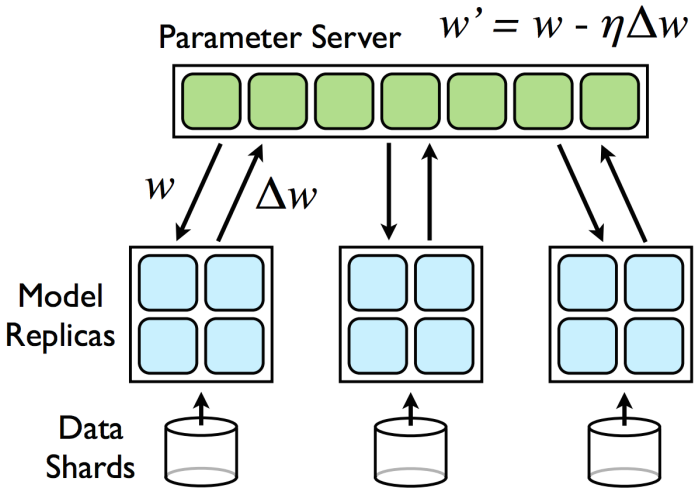

图 2.2 分布式机器学习框架

供给其他学习器，并重复此训练过程。

2.5 大规模核方法

核方法需要存储核矩阵、同时存在矩阵运算，因此核方法的时间、空间复杂度均不低于 $\mathcal{O}(n^2)$，从而导致核方法不适用于大规模问题。目前用于解决核方法大规模瓶颈的方法主要有分布式方法[137]、低秩近似方法[30,31]、随机梯度下降方法[33,138,139] 等三种主流方法。

2.5.1 分布式方法

分布式方法广泛地用于大规模机器学习算法（包括核方法）中，是解决大规模学习算法的存储、计算瓶颈的通用手段。如图 2.2 所示（来源于参数服务器[140]），分布式方法将训练数据划分到不同的计算节点上，计算节点分别对本地数据进行训练局部模型，在训练过程中与中心节点进行必要通信，最后将局部模型汇总到中心节点进行合并为最终模型[141]。

常用的分布式机器学习框架包括 MapReduce[142]、Apache Spark[143]、参数服务器 (parameter sever)[140]、 MLBase[144] 等。已有大量分布式算法研究，包括提高算法收敛率速度[145–147]，降低通信复杂度[140,148,149]。同时，有大量研究将核方法应用到分布式环境中，包括分布式核岭回归[150–152]、分布式支持向量机[153–155]、分布式核主成分分析[156]、分布式谱聚类[157] 等。

而分布式核方法作为近似核方法，存在一定精度损失，分布式核方法的泛

化理论研究成为研究难点与热点。而分治算法框架 (divide and conquer) 是通信最少的分布式框架（只通信一次），同时其精度损失也最多，因此分布式核方法的泛化理论研究通常基于分治算法。 Zhang 等基于积分算子理论、核矩阵特征分解，首次给出分布式核岭回归的最优泛化误差保证[28,137]； Wang 等进一步降低了样本复杂度、得到更实用的 $1+\epsilon$ 界[158]。 Lin 等首次基于传统的积分算子理论，使用容量假设、正则化假设，得到了分布式核岭回归的最优泛化误差收敛率[75]；基于此理论 Chang 等分析了半监督分布式核岭回归学习，并证明了使用无标签数据可以用来增大分块数。而 Meng 等首次使用稳定性 (stability) 对分布式核岭回归的泛化性能进行分析[159]。近年来，将分布式算法与梯度下降算法的结合并分析其最优泛化理论也成为研究热点[79,160]。

2.5.2　低秩近似方法

由于非线性核方法的理论时间复杂度不会低于 $\mathcal{O}(n^2)$，因此研究的重点是设计快速近似算法。低秩近似方法主要包括：使用Nyström 采样近似核矩阵、使用随机特征近似核函数。

2.5.2.1　Nyström 采样

Nyström方法[30] 对训练集采样，使用采样得到的样本子集对原始核矩阵进行低秩近似。使用 s 个 Nyström 采样中心点，Nyström 采样近似核矩阵为

$$\mathbf{K}_{nn} \approx \mathbf{K}_{ns}\mathbf{K}_{ss}^{\dagger}\mathbf{K}_{ns}^{T},$$

其中 $\mathbf{K}_{nn}$ 为完整的核矩阵，$\mathbf{K}_{ns}$ 为定义在全部数据集、采样数据上的核矩阵，$\mathbf{K}_{ss}$ 为定义在采样数据上的核矩阵。$\dagger$ 代表矩阵伪逆。针对 Nyström 方法的研究主要包括以下五个方面

(1) **生成 Nyström 采样点**

Nyström 采样点决定了近似核方法的泛化性能，已有大量工作研究如何高效地生成更合适的 Nyström 采样点，包括使用 k-means 聚类中心作为 Nyström 采样点[161]； Kumar 等提出将多个标准 Nyström 近似进行组合的集成 Nyström 方法[162]； Li 等使用随机策略生成 Nyström 采样点[163]； Hsieh 等提出伪里程碑点 (pseudo landmark points) 作为 Nyström 采样点[164]。同时对 Nyström 采样加速，进一步降低 Nyström 采样的时间空间、复杂度也成为研究热点。 De Brabanter 等使

用基于熵的 Nyström 采样，提高泛化性能、计算效率[165]；Gittens 和 Mahoney 使用了基于打分函数 (leverage score function) 减少了所需的 Nyström 采样点数[166]，从而提高了 Nyström 近似效率。

(2) **基于矩阵操作加速 Nyström 近似**

Si 等提出使用矩阵的快速变换加速 Nyström 近似的 Fast-Nys 方法，提高了计算效率[167]。对核矩阵进行划分后，在分块上使用 Nyström 矩阵，使得只需更少的 Nyström 采样点就能够获得良好的近似效果。该方面研究包括使用聚类进行区块划分后使用分块对角元素 (block diagonal) 降低 Nyström 空间复杂度、提高计算效率[168]；生成树结构后，进行层次划分近似组合核的层次组合核 (hierarchically compositional kernel) 方法[169]。

(3) **Nyström 泛化理论**

在 Nyström 核方法的泛化理论分析方面也取得了进展，Bach 对 Nyström 方法的泛化理论进行分析[170]； Alaoui 和 Mahoney 给出 Nyström 加速的核岭回归的理论保证[171]；而基于积分算子理论框架，Rudi 等证明了 Nyström 近似核岭回归的最优泛化误差率[73]。

(4) **使用随机优化方法加速 Nyström 核方法求解**

近年来，很多研究工作将 Nyström 加速的核岭回归与梯度下降算法相结合，提高计算效率的同时保证最优泛化误差收敛率，包括使用早停 (early stopping) 加速 Nyström–KRR 求解的方法[172]；将 Nyström 近似方法与坐标梯度下降相结合的方法[173]；使用预处理共轭梯度下降算法 (preconditioning conjugate gradient descent, PCG) 加速 Nyström–KRR 闭式解求解的 FALKON 方法[78]。

2.5.2.2 随机特征

随机特征 (random feature) 使用有限维随机特征 $\phi_M : \mathcal{X} \to \mathbb{R}^M$ 近似核函数

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \phi_M(\boldsymbol{x}), \phi_M(\boldsymbol{x}') \rangle.$$

基于傅里叶逆变换，Rahimi 和 Recht 首次为平移不变核提出使用随机傅里叶特征[31]。由于随机特征方法在泛化性能上与核方法相似，而求解上显式特征映射后可以使用线性方法，兼具了良好的泛化性能、较高的计算效率，从而引发了随机特征方法的研究与应用热潮[174–176]。

(1) **其他核函数对应随机特征**

除平移不变核外，研究者们也为其他类型核函数构造对应的随机特征，包括点积核（包括线性核、多项式核等）[177–179]、直方图核 (histogram kernel)[180]、附加内核 (additive kernel)[181]、半群核[182] 等。 Samo 和 Roberts 提出使用傅里叶随机特征近似非平稳谱核[101]，并在高斯过程[102–104] 上得到广泛应用。

(2) **提高随机特征映射效率**

Le 等使用对角高斯矩阵、Hadamard 矩阵近似随机高斯矩阵的方法 (Fastfood)，提高了特征映射的计算速度[175]。 Yang 等使用拟蒙特卡洛 (Quasi-Monte Carlo) 采样的随机傅里叶特征近似平移不变核，加速特征映射求解[176]。Agrawal 等使用 0 范数获得随机特征的稀疏解，从而获得压缩后的低秩近似[183]。

(3) **随机特征泛化理论研究**

Sriperumbudur 和 Szabó 研究了随机傅里叶特征的最优泛化理论[184]。 Rudi 和 Rosasco 基于积分算子理论，为随机特征方法给出通用的最优收敛率证明[74]。

(4) **随机特征与其它加速方法结合的方法**

Avron 等使用预处理加速求解随机特征近似的核岭回归[185]，并给出近似误差分析。 Carratino 等将随机梯度下降 (SGD) 与随机特征相结合[81]，并给出最优泛化理论保证。 McWilliams 等将随机特征与分布式算法相结合用于求解大规模核岭回归问题[186,187]，但没有给出泛化理论分析；而已完成工作基于积分算子理论学习框架，为分布式、随机特征相结合的方法提供了最优泛化理论保证[188]。

2.5.2.3　其他近似方法

其他近似方法如近似SVD分解[189]、核展开[190]、稀疏贪婪矩阵(sparse greedy matrix)[191]、快速高斯变换[192]、树编码(tree code)[193] 等。

2.5.3　随机优化方法

随机优化 (stochastic optimization) 方法利用凸优化技术（主要是一阶梯度下降算法），加速核方法的求解。核方法中常用的随机优化方法包括以下四种

(1) **割平面法 (cutting plane algorithm)**

针对线性核学习器，Joachims 提出了基于割平面法的 SVM^{perf} 具有线性时间复杂度[194]，之后又将割平面法扩展到结构化 SVM (structrual SVM) 中。 Franc 和 Sonnenburg 基于割平面法的设计出更快的线性 SVM 方法 (OCAS)[195]。

(2) **坐标下降法 (coordinate descent)**

基于坐标下降的序列最小化优化方法 (SMO, sequential minimal optimization) 由 Platt 在1998年发明，目前被广泛地应用于 SVM 的训练中[32]，首次避免了使用二次规划方法，并在通用的 SVM 库 LIBSVM 中得以实现[45]。之后，Chang 等又提出了原坐标下降算法 (primal coordinate descent, PCD)[196]、对偶坐标下降算法 (dual coordinate descent)[197]。

(3) **梯度方法 (gradient methods)**

Bottou 和 Bousquet 对常用的一阶梯度下降算法、二阶梯度下降算法、随机梯度下降算法进行分析比较，讨论不同梯度方法的适用情况[198]。

一阶随机梯度下降算法通过优化目标函数的负梯度方向来更新权重

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^{t} - \eta \frac{\partial g(\boldsymbol{W})}{\partial \boldsymbol{W}},$$

其中 η 为学习率，$g(\boldsymbol{W})$ 为优化目标。 Shalev-Shwartz 等提出的 Pegasos 学习框架首次将随机梯度下降方法引入到非线性 SVM 中，并对梯度方法求解 SVM 的收敛率进行了讨论[33]。基于 Pegasos 方法，Orabona 和 Luo 设计了多核学习的随机梯度下降算法 (UFO)[97]。 Dai 等为大规模核方法设计了双重梯度下降算法 (doubly stochastic gradients)[138]。近期梯度方法动态学习率的研究也可以扩展到核方法求解中，包括 Adagrad[199]、Adadelta[200]、Adam[201] 等。

而二阶梯度下降算法，无需定义步长，能够很快收敛，定义如下

$$\boldsymbol{W}^{t+1} = \boldsymbol{W} - \boldsymbol{H}^{-1} \frac{\partial g(\boldsymbol{W})}{\partial \boldsymbol{W}},$$

其中 $\boldsymbol{H}^{-1}$ 是海塞矩阵 (Hessian matrix) 的逆。牛顿法 (Newton method) 是常用的二阶梯度下降算法，有着收敛速度快的优点[202]。但牛顿法在每次迭代中需要求解海塞矩阵的逆，计算复杂，通常使用拟牛顿法 (quasi Newton method) 求解近似的海塞矩阵逆或海塞矩阵，主要包括 DFP、BFGS、Broyden等方法[203]。

(4) **预处理共轭梯度下降方法 (preconditioning conjugate gradient, PCG)**

共轭梯度下降方法 (conjugate gradient, CG) 用于迭代求解线性方程，其迭代次数依赖于线性方程中矩阵的条件数 (condition number)。预处理共轭梯度下降方法 (preconditioning conjugate gradient, PCG) 引入预处理器，降低线性方程中矩阵条件数，从而减少求解的迭代次数。 PCG 方法广泛地应用于加速核岭回归、近似核岭回归的闭式解求解中[78,139,185,204]。

第3章　大规模半监督的核方法模型选择泛化理论研究

对于传统的半监督学习问题，输入空间为 $\mathcal{X} = \mathbb{R}^d$，输出空间为 $\mathcal{Y} \subseteq \mathbb{R}^K$，$\mathcal{X} \times \mathcal{Y}$ 存在一个固定但未知的联合概率密度分布 $\rho_{X\times Y}$。考虑通用的问题设定，当 $K = 1$ 时，输出标签为单变量，如常见的二分类、回归问题；当 $K > 1$ 时，输出标签为多元变量，如多分类、多标签、多任务问题。对于半监督问题，有标签训练集合 $D^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ 中成对的输入、输出样本独立同分布地采样于联合概率密度分布 $\rho_{X\times Y}$，无标签数据集合 $D^u = \{\boldsymbol{x}_i\}_{i=1}^u \in \mathcal{X}$ 中输入样本独立同分布地采样于边际概率分布 ρ_X。在常见的半监督学习、应用中，有标签样本数远远少于无标签样本数，即 $n \ll u$。

核方法是求解半监督问题的常见方法，有坚实的理论基础和完备的学习框架。核方法利用核函数 $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 诱导的隐式特征映射 $\phi : \mathcal{X} \to \mathcal{H}$，将输入样本从输入空间隐式地映射到再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS) $\mathcal{H}$ 中，然后在 $\mathcal{H}$ 中训练线性学习器。核函数是特征映射的内积形式 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}'), \phi(\boldsymbol{x}) \rangle$。相对于传统线性方法直接在输入空间 $\mathcal{X}$ 上学习模型，核方法在通常更高维的隐式特征空间 $\mathcal{H}$ 中学习模型，往往拥有更强大的特征表达能力，带来了泛化性能的提升。

在本章中，首先介绍核方法的泛化理论所需要的假设空间、泛化误差、泛化误差界等定义[107,205]；基于预备知识，再依次介绍基于核矩阵谱分析的谱度量泛化理论、基于核函数谱分析的 Rademacher 复杂度泛化理论、基于积分算子理论的最优泛化理论。

3.1　预备知识

定义 3.1 (假设空间)**.** 核方法通过隐式特征映射 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{H}}$ 将输入样本映射到在再生核希尔伯特空间 $\mathcal{H}$，之后在 $\mathcal{H}$ 上训练线性学习器 $f(\boldsymbol{x}) = \langle \boldsymbol{W}, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}}$，对应的假设空间 H_κ 定义为

$$H_\kappa = \left\{ f \mid \boldsymbol{x} \to f(\boldsymbol{x}) = \langle \boldsymbol{W}, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}} : \|\boldsymbol{W}\|_p \leq 1 \right\}, \tag{3.1}$$

其中，$\boldsymbol{W} \in \mathcal{H} \times \mathcal{Y}$ 为核学习器对应的权重矩阵，$\|\boldsymbol{W}\|_p$ 为定义在 $\mathcal{H}$ 上的范数用

于正则化假设空间。$\phi : \mathcal{X} \to \mathcal{H}$ 将输入样本 $\boldsymbol{x}$ 从输入空间 $\mathcal{X}$ 映射到再生核希尔伯特空间 $\mathcal{H}$，进而通过线性学习器 $\langle \boldsymbol{W}, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}}$ 映射到输出空间 $\mathcal{Y}$。

基于上述假设空间定义，（半）监督学习的目标就是最小化学习模型的泛化误差 (generalization error)，也叫期望误差 (expected error)，从而最大化学习器的泛化性能（对未知数据的预测能力）。

定义 3.2. 常见损失函数包括：分类问题使用的合页损失 (hinge loss)

$$\ell(f(\boldsymbol{x}), \boldsymbol{y}) = \left| 1 - \left(\boldsymbol{y}^\top f(\boldsymbol{x}) - \max_{\boldsymbol{y}' \neq \boldsymbol{y}} \boldsymbol{y}'^\top f(\boldsymbol{x}) \right) \right|_+ .$$

以及回归问题使用的平方损失 (squared loss)：

$$\ell(f(\boldsymbol{x}), \boldsymbol{y}) = \| f(\boldsymbol{x}) - \boldsymbol{y} \|_2^2 .$$

定义 3.3 (最小化泛化误差)**.** 半监督核方法的学习目标为最小化模型泛化误差

$$\inf_{f \in \boldsymbol{H}_\kappa}, \quad \mathcal{E}(f) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f(\boldsymbol{x}), \boldsymbol{y}) \, d \, \rho_{X \times Y}(\boldsymbol{x}, \boldsymbol{y}),$$

其中，$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 为定义在输出空间上的损失函数。

由于潜在的联合概率分布 $\rho_{X \times Y}$ 未知，无法直接计算模型的泛化误差。实际上，统计学习理论中模型的学习目标为：最小化经验学习器的泛化误差与假设空间中最优学习器的泛化误差的差异（泛化误差上界）。

定义 3.4 (最小化经验误差)**.** 令 $\widehat{f_n}$ 为假设空间中经验损失最小化假设 (empirical risk minimization, ERM)。基于有标签训练数据 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$，经验模型使用经验损失最小化（ERM）作为学习目标进行训练

$$\widehat{f_n} := \arg\min_{f \in \boldsymbol{H}_\kappa} \widehat{\mathcal{E}}(f) = \arg\min_{\boldsymbol{W} \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i) + \lambda_A \| \boldsymbol{W} \|_p . \tag{3.2}$$

其中 $\| \boldsymbol{W} \|_p$ 为正则化项、λ_A 为正则化项参数。

定义 3.5 (最小化泛化误差界)**.** 统计学习理论中，学习目标为最小化泛化误差界

$$\inf_{\widehat{f_n} \in \boldsymbol{H}_\kappa}, \ \mathcal{E}(\widehat{f_n}) - \mathcal{E}(f^*), \tag{3.3}$$

其中，基于训练数据集实际学得的模型 $\widehat{f_n}$ 为假设空间中经验损失最小化假设 (empirical risk minimization, ERM)。令假设空间中中存在最优模型（即期望损失最小的模型） f^*，即 $\min_{f \in \boldsymbol{H}_\kappa} \mathcal{E}(f) = \inf_{f \in \boldsymbol{H}_\kappa} \mathcal{E}(f)$。一般机器学习任务（如二分类、回归任务）的泛化误差收敛率为 $\mathcal{O}(1/\sqrt{n})$。

最优泛化性能对应于最小泛化误差，但泛化误差无法直接进行计算。如定义 3.5 所示，泛化理论分析通常通过最小化泛化误差界获得更好的泛化性能。本章使用谱度量、Rademacher 复杂度、积分算子理论等工具，界定半监督核方法泛化误差上界。通过最小化最小泛化误差上界，指导模型选择准则、大规模半监督算法的设计。已完成工作对下面三种不同的泛化误差界进行了学习：

(1) 基于核矩阵谱分析的谱度量泛化误差理论

通过核矩阵谱分析，定义出谱度量 (spectral measure)，并使用谱度量建立二分类核方法的泛化误差上界[206]。

(2) 基于核函数谱分析的 Rademacher 复杂度泛化误差理论

首次将局部 Rademacher 复杂度引入到多分类核学习器中，获得线性依赖于样本数的泛化误差界[205]。针对线性多分类，建立定义在损失空间上 Rademacher 复杂度与假设空间上 Rademacher 复杂度的关系，使用权重矩阵的尾部奇异值之和界定线性多分类的局部 Rademacher 复杂度，建立基于局部 Rademacher 复杂度的线性多分类泛化误差界[207]。针对多输出问题，给出通用的基于局部 Rademacher 复杂度的半监督多输出核方法泛化误差界，分别界定核学习器对应局部 Rademacher 复杂度、线性学习器对应的局部 Rademahcer 复杂度建立对应的核学习器泛化误差界、线性学习器泛化误差界[208]。

其中，多分类问题基于局部 Rademacher 复杂度的泛化误差界是多输出问题泛化误差界的特例，因此本文直接介绍针对多输出问题核学习器、线性学习器的局部 Rademacher 复杂度泛化误差界。

(3) 基于积分算子理论的最优泛化理论

基于积分算子理论，使用有效维 (effective dimension) 度量假设空间大小，引入容量假设、正则化假设，进而为结合分布式、随机特征的核岭回归方法推导出最优泛化收敛率[188]。基于积分算子理论，为结合 Nyström 采样、PCG 的半监督核岭回归 (LapRLS) 给出泛化理论保证[135]。

3.2　基于谱度量的二分类泛化误差界

本节使用核矩阵的谱分解手段，给出核矩阵谱度量定义，推导最小二乘支持向量机 (LSSVM)、支持向量机 (SVM) 对应的泛化误差界。对于有监督的二分类问题，其输出空间为 $\mathcal{Y} = \{+1, -1\}$，使用合页损失 $\ell(f(\boldsymbol{x}), \boldsymbol{y}) = |1 - \boldsymbol{y} f(\boldsymbol{x})|_+$。

0-1 损失的二分类泛化误差定义为

$$\mathcal{E}(f) = \Pr_{\rho_{X\times Y}}[yf(\boldsymbol{x}) < 0].$$

二分类经验误差定义为

$$\widehat{\mathcal{E}}(f) = \Pr_{D^l}[yf(\boldsymbol{x}) \le 0].$$

核矩阵定义为 $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^n$，对应的正则化核矩阵为 $\mathbf{N} = \mathbf{K}/|\mathbf{K}|_1$，正则化系数为 $|\mathbf{K}|_1 = \sum_{i,j=1}^n \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$。令 $(\lambda_i, \mathbf{v}_i)$ 为规范化核矩阵 $\mathbf{N}$ 的谱分解，

$$\mathbf{N}\mathbf{v_i} = \lambda_i \mathbf{v_i},$$

其中，$i = 1, \ldots, n$。 λ_i 为降序排列的特征值，$\mathbf{v}_i$ 为对应的特征向量。为方便推导，令正则化参数为常数 $|\mathbf{K}|_1 = C_{|\mathbf{K}|}$。对于任意两个输入样本 $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$，核矩阵中元素存在上界 $0 \le \kappa(\boldsymbol{x}, \boldsymbol{x}') \le c_\kappa$。

3.2.1 谱度量定义

核方法的泛化性能由核函数决定，而核矩阵中包含了对应核函数的信息，反映了核函数与具体任务（由数据分布 $\rho_{X\times Y}$ 决定）的匹配程度。而核矩阵谱分解是理解核矩阵的重要手段，本节基于核矩阵谱分解给出谱度量，从而能够将谱分解理论应用于核方法模型选择中。

定义 3.6 (谱度量 (spectral measure, SM))**.** 令 $(\lambda_i, \mathbf{v}_i)$ 为正则化核矩阵 $\mathbf{N}$ 的谱分解，$i = 1, \ldots, n$。假定 $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ 为正实数上的函数，对于任意 $i \in \{1, \ldots, n\}$，存在 $\varphi(\lambda_i) \le \lambda_i$。核矩阵 K 的谱度量定义为

$$\mathrm{SM}(\kappa, \varphi) := \frac{1}{n}\sum_{i=1}^n \varphi(\lambda_i)\langle \boldsymbol{y}, \mathbf{v}_i\rangle^2,$$

其中，$\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^{\mathrm{T}}$。

接下来，给出满足条件 $\varphi(\lambda_i) \le \lambda_i$ 的两种形式:

- 合页形式: $h \ge 0$

$$\varphi(\lambda_i) = \begin{cases} 0 & \text{if } \lambda_i \le h, \\ \lambda_i & \text{其他}. \end{cases}$$

- 高阶形式:

$$\varphi(\lambda_i) = \lambda_i^r, r \geq 1.$$

对于合页形式，存在截断值 h，可以很容易地验证 $\varphi(\lambda_i) \leq \lambda_i$；对于高阶形式，由正则化核矩阵 $\mathbf{N}$ 的定义可知 $0 \leq \lambda_i \leq 1$，所以高阶形式也满足 $\varphi(\lambda_i) \leq \lambda_i$。

当特征值 λ_i 非常小的时候，合页形式为 $\varphi(\lambda_i) = 0$，而高阶形式趋近于0，即 $\varphi(\lambda_i) \to 0$。核矩阵对应的小特征值由噪声造成[209]，而函数 $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ 的使用能够达到去除噪声的作用。

3.2.2　LSSVM 的谱度量泛化误差界

基于定义 3.4 中经验风险最小化定义 (3.2)，最小二乘支持向量机 (LSSVM) 使用平方损失 $\ell(f(\boldsymbol{x}), \boldsymbol{y}) = (f(\boldsymbol{x}) - \boldsymbol{y})^2$，经验误差最小化学习器为:

$$\widehat{f_n} = \arg\min_{f \in H_\kappa} \left\{ \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i) + \lambda_A \|\boldsymbol{W}\|_{\mathcal{H}}^2 \right\}, \tag{3.4}$$

在推导 LSSVM 的谱度量泛化误差界之前，首先推导出引理 3.1 给出经验误差与谱度量的关系。接下来再介绍由 Gao 等给出的定理[210]，该定理讨论了泛化误差与经验误差的关系。

引理 3.1. *给定有标签数据集 $D^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$，对于 LSSVM 假设空间中任意学习器 $f \in H_\kappa$，存在*

$$\frac{1}{n} \sum_{i=1}^{n} y_i f(\boldsymbol{x}_i) \geq c_0 \cdot \mathrm{SM}(\kappa, \varphi). \tag{3.5}$$

证明. 由LSSVM学习器(3.4)、平方损失，LSSVM存在闭式解 $(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^{\mathrm{T}} = \mathbf{K}\boldsymbol{\alpha}$. 其中，$\boldsymbol{\alpha} = [\mathbf{K} + \lambda_A \mathbf{I}]^{-1} \boldsymbol{y}$。因此有

$$\sum_{i=1}^{n} \boldsymbol{y}_i f(\boldsymbol{x}_i) = \boldsymbol{y}^{\mathrm{T}} \mathbf{K}\boldsymbol{\alpha} = \boldsymbol{y}^{\mathrm{T}} \mathbf{K}[\mathbf{K} + \lambda_A \mathbf{I}]^{-1} \boldsymbol{y}. \tag{3.6}$$

令 $(\beta_i, \mathbf{v}_i)$ 为核矩阵 $\mathbf{K}$的谱分解，则

$$\boldsymbol{y}^{\mathrm{T}} \mathbf{K}[\mathbf{K} + \lambda_A \mathbf{I}]^{-1} \boldsymbol{y} = \sum_{i=1}^{n} \frac{\beta_i}{\lambda_A + \beta_i} \langle \boldsymbol{y}, \mathbf{v}_i \rangle^2. \tag{3.7}$$

由 $\mathrm{Tr}(\mathbf{K}) \leq |\mathbf{K}|_1 = C_{|\mathbf{K}|}$，可得

$$\frac{\beta_j}{C_{|\mathbf{K}|}} \leq \frac{\mathrm{Tr}(\mathbf{K})}{C_{|\mathbf{K}|}} \leq 1.$$

根据等式 (3.7)，可得

$$\boldsymbol{y}^{\mathrm{T}}\mathbf{K}[\mathbf{K}+\lambda_A\mathbf{I}]^{-1}\boldsymbol{y}=\sum_{i=1}^{n}\frac{\beta_i/C_{|\mathbf{K}|}}{\beta_i/C_{|\mathbf{K}|}+\lambda_A/C_{|\mathbf{K}|}}\langle\boldsymbol{y},\mathbf{v}_i\rangle^2\geq\sum_{i=1}^{n}\frac{\beta_i/C_{|\mathbf{K}|}}{1+\lambda_A/C_{|\mathbf{K}|}}\langle\boldsymbol{y},\mathbf{v}_i\rangle^2.$$

规范化核矩阵 $\mathbf{N}$ 对应特征值为 $\lambda_i=\beta_i/|\mathbf{K}|_1=\beta_i/C_{|\mathbf{K}|}$，特征向量为 $\mathbf{v}_i$。故有

$$\boldsymbol{y}^{\mathrm{T}}\mathbf{K}[\mathbf{K}+\lambda_A\mathbf{I}]^{-1}\boldsymbol{y}\geq\sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_A/C_{|\mathbf{K}|}}\langle\boldsymbol{y},\mathbf{v}_i\rangle^2. \tag{3.8}$$

由函数 $\varphi(\lambda_i)\leq\lambda_i$，结合 (3.6)、(3.7)、(3.8)，可得

$$\sum_{i=1}^{n}\boldsymbol{y}_i f(\boldsymbol{x}_i)\geq\sum_{i=1}^{n}\frac{\varphi(\lambda_i)}{1+\lambda_A/C_{|\mathbf{K}|}}\langle\boldsymbol{y},\mathbf{v}_i\rangle^2=c_0\cdot\mathrm{SM}(\kappa,\varphi).$$

其中，$c_0=\frac{C_{|\mathbf{K}|}}{n(C_{|\mathbf{K}|}+\lambda_A)}$。 □

引理 3.2 (定理 8[210]). 有标签数据集 $D^l=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$ 独立同分布采样于分布 $\rho_{X\times Y}$，且样本个数 $n\geq 5$。对于任意 $f\in H_\kappa,\delta\in(0,1)$，以至少 $1-\delta$ 的概率存在

$$\mathcal{E}(f)\ \leq\frac{2}{n}+\inf_{\theta\in(0,1]}\left[\Pr_{D^l}[\boldsymbol{y}f(\boldsymbol{x})\leq\theta]+\quad\frac{7\mu+3\sqrt{3\mu}}{3n}+\sqrt{\frac{3\mu}{n}\Pr_{D^l}[\boldsymbol{y}f(\boldsymbol{x})\leq\theta]}\right],$$

其中，$\mu=\frac{8}{\theta^2}\ln n\ln(2n)+\ln\frac{2n}{\delta}$。

定理 3.3. 假设 $\|\boldsymbol{W}\|_{\mathcal{H}}\leq 1$，假设空间中最优模型泛化误差能够取到 $\inf_{f\in H_\kappa}\mathcal{E}(f)=0$。*LSSVM*的经验模型 $\widehat{f_n}$ 由 (3.4) 给出。对于随机采样的训练集 $D^l=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)_{i=1}^n\}$，其训练样本个数 $n\geq 5$，以至少 $1-\delta$ 的概率存在如下泛化误差界

$$\mathcal{E}(\widehat{f_n})-\mathcal{E}(f^*)\leq 1-c_0\cdot\mathrm{SM}(\kappa,\varphi)+\inf_{\theta\in(0,1]}\left[\theta+\frac{7\mu+3\sqrt{3\mu}+6}{3n}+\sqrt{\frac{3\mu}{n}}\right],$$

其中，$\mu=\frac{8}{\theta^2}\ln n\ln(2n)+\ln\frac{2n}{\delta}$，$c_0=\frac{C_{|\mathbf{K}|}\lambda_A}{C_{|\mathbf{K}|}+\lambda_A}$。

证明. 松弛后的经验误差定义为

$$\Pr_{D^l}[y\widehat{f_n}(\boldsymbol{x})\leq\theta]=\frac{1}{n}\sum_{i=1}^{n}1_{[\boldsymbol{y}_i\widehat{f_n}(\boldsymbol{x}_i)-\theta]},$$

如果 $t<0$ 则 $1_{[t]}=1$，否则 $1_{[t]}=0$。易得

$$\Pr_{D^l}[y\widehat{f_n}(\boldsymbol{x})\leq\theta]\leq\frac{1}{n}\sum_{i=1}^{n}\max\left\{0,1-\boldsymbol{y}_i\widehat{f_n}(\boldsymbol{x}_i)+\theta\right\}. \tag{3.9}$$

从假设空间的定义可知 $1-\boldsymbol{y}_i\widehat{f_n}(\boldsymbol{x}_i)+\theta>0$。因此，基于等式 (3.9)，可得

$$\Pr_{D^l}[y\widehat{f_n}(\boldsymbol{x})\leq\theta]\leq\frac{1}{n}\sum_{i=1}^{n}(1-\boldsymbol{y}_i\widehat{f_n}(\boldsymbol{x}_i)+\theta). \tag{3.10}$$

根据等式 (3.10)、等式 (3.5)，易得

$$\Pr_{D^l}[y\widehat{f}_n(\boldsymbol{x}) \le \theta] \le 1 + \theta - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \widehat{f}_n(\boldsymbol{x}_i) \le 1 + \theta - c_0 \cdot \mathrm{SM}(\kappa, \varphi). \tag{3.11}$$

将等式 (3.11) 带入引理 3.2中，可证得

$$\Pr_{D^l}[y\widehat{f}_n(\boldsymbol{x}) < 0] \le \frac{2}{n} + \inf_{\theta\in(0,1]} \Big[[1 + \theta - c_0 \cdot \mathrm{SM}(\kappa, \varphi)] + \frac{7\mu + 3\sqrt{3\mu}}{3n} + \sqrt{\frac{3\mu}{n}} \Big].$$

又由经验误差定义为 $\mathcal{E}(\widehat{f}_n) = \Pr_{D^l}[y\widehat{f}_n(\boldsymbol{x}) < 0]$，而最小化泛化误差 $\mathcal{E}(f^*) = \inf_{f\in H_\kappa} \mathcal{E}(f) = 0$，可证得定理。 □

3.2.3　SVM 的谱度量泛化误差界

基于定义 3.1中假设空间 H_κ， SVM 使用合页损失 $\ell(f(\boldsymbol{x}), \boldsymbol{y}) = |1 - \boldsymbol{y} f(\boldsymbol{x})|_+$，经验最小化 (ERM) 学习器定义为 (3.4)。类似于 LSSVM 中的推导，在推导 SVM 的谱度量泛化误差界之前，首先推导出引理 3.1 给出经验误差与谱度量的关系。

引理 3.4. *给定有标签数据集 $D^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$，对于 SVM 假设空间中任意学习器 $f \in H_\kappa$，存在如下不等式*

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i f(\boldsymbol{x}_i) \ge C_{|\mathbf{K}|} \left(\mathrm{SM}(\kappa, \varphi) + \frac{C_\lambda}{n} \right),$$

其中，$C_\lambda = \min\{-1, 1 - \frac{1}{2\lambda_A}\}$。

证明. 由表示定理可得 $f(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x})$, 其中，$\boldsymbol{\alpha}$ 是 SVM 的对偶解。可得

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i f(\boldsymbol{x}_i) = \frac{1}{n}\boldsymbol{y}^{\mathrm{T}} \kappa(\boldsymbol{y} \otimes \boldsymbol{\alpha}), \tag{3.12}$$

其中，$\otimes$ 为矩阵中相同位置元素点对点乘积（称为 Hadamard 乘积)。易得

$$\begin{aligned}
&\boldsymbol{y}^{\mathrm{T}} \kappa(\boldsymbol{y} \otimes \boldsymbol{\alpha}) - \boldsymbol{y}^{\mathrm{T}} \mathbf{K} \boldsymbol{y} \\
&= \Big[\sum_{\boldsymbol{y}_i = \boldsymbol{y}_j} \alpha_i \mathbf{K}_{ij} - \sum_{\boldsymbol{y}_i \ne \boldsymbol{y}_j} \alpha_i \mathbf{K}_{ij} \Big] - \Big[\sum_{\boldsymbol{y}_i = \boldsymbol{y}_j} \mathbf{K}_{ij} - \sum_{\boldsymbol{y}_i \ne \boldsymbol{y}_j} \mathbf{K}_{ij} \Big] \\
&= \Big[\sum_{\boldsymbol{y}_i = \boldsymbol{y}_j} \alpha_i \mathbf{K}_{ij} + \sum_{\boldsymbol{y}_i \ne \boldsymbol{y}_j} \mathbf{K}_{ij} \Big] - \Big[\sum_{\boldsymbol{y}_i \ne \boldsymbol{y}_j} \alpha_i \mathbf{K}_{ij} + \sum_{\boldsymbol{y}_i = \boldsymbol{y}_j} \mathbf{K}_{ij} \Big] \\
&\ge \sum_{i,j} \min\{\alpha_i, 1\} \cdot \mathbf{K}_{ij} - \sum_{i,j} \max\{\alpha_i, 1\} \cdot \mathbf{K}_{ij} \\
&= \sum_{i,j} (\min\{\alpha_i, 1\} - \max\{\alpha_i, 1\}) \cdot \mathbf{K}_{ij} \\
&= \sum_{i,j} c_i \mathbf{K}_{ij},
\end{aligned} \tag{3.13}$$

其中，$c_i = \min\{\alpha_i, 1\} - \max\{\alpha_i, 1\}$。由 $0 \le \alpha_i \le \frac{1}{2\lambda_A}$，可得

$$c_i = \min\{\alpha_i - 1, 1 - \alpha_i\} \ge \min\{-1, 1 - \frac{1}{2\lambda_A}\} =: C_\lambda.$$

由等式 (3.13)，可得

$$\boldsymbol{y}^{\mathrm{T}}\kappa(\boldsymbol{y} \otimes \boldsymbol{\alpha}) - \boldsymbol{y}^{\mathrm{T}}\mathbf{K}\boldsymbol{y} \ge C_\lambda \sum_{i,j} \mathbf{K}_{ij} = C_\lambda \cdot C_{|\mathbf{K}|}.$$

因此可得

$$\boldsymbol{y}^{\mathrm{T}}\kappa(\boldsymbol{y} \otimes \boldsymbol{\alpha}) \ge dC + \boldsymbol{y}^{\mathrm{T}}\mathbf{K}\boldsymbol{y} = C_\lambda C_{|\mathbf{K}|} + \sum_i \beta_i \langle \boldsymbol{y}_i, \mathbf{u}_i \rangle^2 =$$

$$C_\lambda C_{|\mathbf{K}|} + C_{|\mathbf{K}|} \sum_i \lambda_i \langle \boldsymbol{y}_i, \mathbf{v}_i \rangle^2 \ge C_\lambda C_{|\mathbf{K}|} + C_{|\mathbf{K}|} \sum_i \varphi(\lambda_i) \cdot \langle \boldsymbol{y}_i, \mathbf{v}_i \rangle^2,$$

其中，$(\beta_i, \mathbf{u}_i)$ 为核矩阵 $\mathbf{K}$ 的谱分解。引理得证。 □

定理 3.5. 假设 $\|\boldsymbol{W}\|_{\mathcal{H}} \le 1$，假设空间中最优模型泛化误差能够取到 $\inf_{f \in H_\kappa} \mathcal{E}(f) = 0$。*SVM* 的经验模型 $\widehat{f_n}$ 由 (3.4) 给出。对于随机采样的训练集 $D^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^n\}$，其训练样本个数 $n \ge 5$，以至少 $1 - \delta$ 的概率存在如下泛化误差界

$$\mathcal{E}(\widehat{f_n}) - \mathcal{E}(f^*) \le 1 - C_{|\mathbf{K}|} \cdot \mathrm{SM}(\kappa, \varphi) +$$
$$\inf_{\theta \in (0,1]} \left[\theta + \frac{7\mu + 3\sqrt{3\mu} + 3/(2\lambda_A) - 3C_\lambda}{3n} + \sqrt{\frac{3\mu}{n}} \right],$$

其中，$\mu = \frac{8}{\theta^2} \ln n \ln(2n) + \ln \frac{2n}{\delta}$，$C_\lambda = \min\{-1, 1 - \frac{1}{2\lambda_A}\}$。

证明. 结合引理 3.4、等式 (3.10)，可得

$$\Pr_{D^l}[y\widehat{f_n}(\boldsymbol{x}) \le \theta] \le 1 + \theta - C_{|\mathbf{K}|} \cdot \mathrm{SM}(\kappa, \varphi) - \frac{C_\lambda C_{|\mathbf{K}|}}{n}. \tag{3.14}$$

经验误差定义为 $\mathcal{E}(\widehat{f_n}) = \Pr_{D^l}[y\widehat{f_n}(\boldsymbol{x}) < 0]$，而最小化泛化误差 $\mathcal{E}(f^*) = \inf_{f \in H_\kappa} \mathcal{E}(f) = 0$。将等式 (3.14) 带入引理 3.2 中，定理得证。 □

3.3 基于谱分析的 Rademacher 复杂度泛化理论

假设空间复杂度是统计学习理论中进行泛化理论分析的关键工具，包括 VC 维、Rademacher 复杂度、覆盖数等假设空间复杂度度量。由于VC维与给定数据分布无关 (distribution free)、数据独立 (data independent) 的，VC维在估计模型泛

化误差界上通常过于保守。 Koltchinskii、Bartelett 等将数据依赖 (data dependent) 的 Rademacher 复杂度引入到泛化理论研究中，获得了更紧致的泛化误差界及更快的收敛速度[49,211]。 Bartelett 等的进一步研究发现，对模型泛化性能起关键作用的不是定义在整个假设空间上的 Rademacher 复杂度，而是由较小方差的函数构成的假设空间的子空间，从而提出了局部 Rademacher 复杂度的概念[65]。

针对通用核方法假设空间定义的多输出问题（定义 3.1），本节使用局部 Rademacher 复杂度对半监督方法使用的核学习器、线性学习器的泛化误差界进行分析，并使用无标签数据提升核方法、线性方法的泛化性能。

多输出学习器预测函数产生 K 个输出，该函数为 $f:\mathbb{R}^d\to\mathbb{R}^K$。基于定义 3.1 中假设空间 (3.1)，在泛化理论分析中，核方法使用 $\ell_{2,1}$ 范数，即 $\|\boldsymbol{W}\|_{2,1}\leq 1$；线性方法使用迹范数 $\|\boldsymbol{W}\|_*\leq 1$。为方便推导，使用有界的损失函数 $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,C_\ell]$，其中 $C_\ell>0$ 为常数。同时，通过正则化手段对核函数的特征映射进行限制 $\langle\boldsymbol{\phi}(\boldsymbol{x}),\boldsymbol{\phi}(\boldsymbol{x})\rangle\leq 1$，从而对于核学习器有 $\sup_{\boldsymbol{x}\in\mathcal{X}}\kappa(\boldsymbol{x},\boldsymbol{x})\leq 1$，对于近似核学习器有 $\mathbb{E}\left[\boldsymbol{\phi}_M(\boldsymbol{x})^\top\boldsymbol{\phi}_M(\boldsymbol{x})\right]\leq 1$，对于线性学习器有 $\mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{x}\right]\leq 1$。

核学习器 令 $\kappa:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 为 Mercer 核，核函数为特征映射的内积形式 $\kappa(\boldsymbol{x},\boldsymbol{x}')=\langle\boldsymbol{\phi}(\boldsymbol{x}),\boldsymbol{\phi}(\boldsymbol{x}')\rangle_{\mathcal{H}}$，对应特征映射为 $\boldsymbol{\phi}:\mathbb{R}^d\to\mathcal{H}$。核学习器为 $f(\boldsymbol{x})=\langle\boldsymbol{W},\boldsymbol{\phi}(\boldsymbol{x})\rangle_{\mathcal{H}},\ \forall\boldsymbol{x}\in\mathcal{X}$，其中 $\boldsymbol{W}\in\mathcal{H}\times\mathcal{Y}$。对于大规模数据，核方法存在计算效率、存储需求的瓶颈，其时间复杂度、空间复杂度均不小于 $\mathcal{O}(n^2)$。

近似核学习器 Rahimi 和 Recht 在2007年[31] 提出了使用随机傅里叶特征来近似平移不变核（包括高斯核、拉普拉斯核等），并得到进一步发展[175,176]。除平移不变核以外，已有研究使用随机特征近似其他类型核函数，包括点积核（包括线性核、多项式核等）[177–179]、半群核[182]等。使用随机特征近似核函数可以形式化地写做 $\kappa(\boldsymbol{x},\boldsymbol{x}')\approx\langle\boldsymbol{\phi}_M(\boldsymbol{x}),\boldsymbol{\phi}_M(\boldsymbol{x}')\rangle$，其中 $\boldsymbol{\phi}(\cdot)$ 显式的特征映射 $\boldsymbol{\phi}_M:\mathbb{R}^d\to\mathbb{R}^M$。近似核学习器可以写为 $f(\boldsymbol{x})=\boldsymbol{W}^\top\boldsymbol{\phi}_M(\boldsymbol{x}),\forall\boldsymbol{x}\in\mathcal{X}$，其中 $\boldsymbol{W}\in\mathbb{R}^{M\times K}$。通过随机特征近似核函数得到的学习器的求解优化效率与线性学习器类似，而实际泛化性能与原始核方法类似。

线性学习器 不使用特征映射，直接在输入空间 $\mathcal{X}=\mathbb{R}^d$ 上进行学习。线性学习器可以写为 $f(\boldsymbol{x})=\boldsymbol{W}^\top\boldsymbol{x},\forall\boldsymbol{x}\in\mathcal{X}$，其中 $\boldsymbol{W}\in\mathbb{R}^{d\times K}$。

本节首先建立定义在损失空间上局部 Rademacher 复杂度与定义在假设空间上局部 Rademacher 复杂度的关联。并使用定义在假设空间上局部 Rademacher

复杂度，为半监督核方法推导出通用的泛化误差界。通过核函数的特征值分解，界定半监督核学习器的局部 Rademacher 复杂度，给出半监督核学习器的泛化误差界；通过线性权重的特征值分解，界定半监督线性学习器的局部 Rademacher 复杂度，给出半监督线性学习器的泛化误差界。

3.3.1 Rademacher 复杂度定义

基于假设空间 H_κ，首先给出损失空间的定义，再分别介绍定义在损失空间上的全局 Rademacher 复杂度、局部 Rademacher 复杂度。假设空间 H_κ 对应损失空间定义为

$$\mathcal{L} = \left\{ \ell(f(\boldsymbol{x}),\ \boldsymbol{y}) \,\middle|\, f \in H_\kappa \right\}. \tag{3.15}$$

定义 3.7 (损失空间上的 Rademacher 复杂度)**.** 损失空间 $\mathcal{L}$ 定义为等式 (3.15)的形式。基于有标签数据集 D^l，损失空间 $\mathcal{L}$ 的经验 Rademacher 复杂度定义为:

$$\widehat{\mathcal{R}}(\mathcal{L}) = \frac{1}{n}\,\mathbb{E}_\epsilon \left[\sup_{\ell \in \mathcal{L}} \sum_{i=1}^{n} \epsilon_i \ell(f(\boldsymbol{x}_i),\ \boldsymbol{y}_i) \right], \tag{3.16}$$

其中，ϵ_i 为相互独立的 Rademacher 随机变量，以相同概率取得 $+1$ 与 -1。其对应的期望 Rademacher 复杂度为 $\mathcal{R}(\mathcal{L}) = \mathbb{E}\,\widehat{\mathcal{R}}(\mathcal{L})$。

定义 3.8 (损失空间上的局部 Rademacher 复杂度)**.** 对于任意 $r > 0$，损失空间的局部 Rademacher 复杂度 $\mathcal{L}$ 定义为

$$\mathcal{R}(\mathcal{L}_r) = \mathcal{R}\left(\left\{ \ell_f \,\middle|\, \ell_f \in \mathcal{L},\ \mathbb{E}\,(\ell_f - \ell_{f^*})^2 \le r \right\} \right), \tag{3.17}$$

其中，ℓ_{f^*} 对应于假设空间中最优学习器最小的泛化误差。$\mathbb{E}\,(\ell_f - \ell_{f^*})^2 \le r$ 界定出与最优误差相近的假设空间子空间。

从全局 Rademacher 复杂度 (3.16) 到局部 Rademacher 复杂度 (3.17)，围绕最小泛化误差 ℓ_{f^*} 以固定半径选取球形子空间 $\mathcal{L}_r \subseteq \mathcal{L}$，对应的假设空间子空间为

$$H_r = \{ f \,|\, f \in H_k,\ \mathbb{E}\,(\ell_f - \ell_{f^*})^2 \le r \}. \tag{3.18}$$

从定义 3.7 可以看出定义在损失空间上的 Rademacher 复杂度依赖于样本标签 $\{\boldsymbol{y}_i\}_{i=1}^n$，因此定义在损失空间上的 Rademacher 复杂度是标签相关的。接下来介绍定义在假设空间上的 Rademacher 复杂度，该复杂度与样本标签无关，在有标签数据集 D^l 和无标签数据集 D^u 上均可以进行估计。

定义 3.9 (假设空间上的局部 Rademacher 复杂度)**.** 令局部假设空间 H_r 定义为 (3.18) 形式。基于有标签数据集 D^l、无标签数据集 D^u，假设空间上的局部 Rademacher 复杂度定义为：

$$\widehat{\mathcal{R}}(H_r) = \frac{1}{n+u}\,\mathbb{E}_{\epsilon}\left[\sup_{f\in H_r}\sum_{i=1}^{n+u}\sum_{k=1}^{K}\boldsymbol{\epsilon}_{ik}\,f_k(\boldsymbol{x}_i)\right],$$

其中 $f_k(\boldsymbol{x}_i)$ 为预测函数 $f(\boldsymbol{x}_i)$ 产生 K 维输出向量中的第 k 个值，$\boldsymbol{\epsilon}_{ik}$ 为 $(n+u)\times K$ 个的 Rademacher 随机变量。对应的期望 Rademacher 复杂度为 $\mathcal{R}(H_r) = \mathbb{E}\,\widehat{\mathcal{R}}(H_r)$。

3.3.2　通用的半监督局部 Rademacher 复杂度泛化误差界

基于多变量输出的 Rademacher 复杂度泛化误差界，需要使用如下假设

假设 3.6 (L-Lipschitz 连续)**.** *对于 $\mathbb{R}^K$ 变量的 ℓ_2 范数，假设多输出学习器的损失函数 $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_+$ 满足 L-Lipschitz 连续*

$$|\ell(f(\boldsymbol{x}),\ \boldsymbol{y}) - \ell(f'(\boldsymbol{x}'),\ \boldsymbol{y})| \le L\|f(\boldsymbol{x}) - f'(\boldsymbol{x}')\|_2,$$

其中 $(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}\times\mathcal{Y}$，任意输入 $\boldsymbol{x}'\in\mathcal{X}$，预测函数 $f, f'\in H_\kappa : \mathcal{X}\to\mathcal{Y}$。

损失函数 ℓ 的 Lipschitz 连续假设是凸损失函数的正则化算法的通用假设。假设 3.6 是多变量输出问题的常见假设，可以推广到结构化预测[212]。利用 Lipschitz 条件和 Rademacher 复杂度的收缩引理 (contraction lemma)[212,213]，从而进一步建立了损失空间上局部 Rademacher 复杂度和假设空间上局部 Rademacher 复杂度之间的关联。

引理 3.7 (引理 5[212])**.** *令损失函数 ℓ 满足假设 3.6，则存在如下收缩不等式*

$$\mathcal{R}(\mathcal{L}_r) \le \sqrt{2}L\mathcal{R}(H_r).$$

Cortes 等[212] 和 Maurer[213] 分别给出了上述引理的证明，在证明过程中使用了 Khintchine 不等式。引理 3.7 中的收缩不等式是分析多变量输出问题泛化性能力的关键工具，连接了定义在损失空间上的 Rademacher 复杂度和定义在假设空间上的 Rademacher 复杂度。假设空间上的 Rademacher 复杂度 $\mathcal{R}(H_r)$ 不依赖于标签，因此将该输出无关的复杂度分析半监督核方法的泛化误差界。

引理 3.8 (定理 3.3[65])**.** *令 $\mathcal{Z}$ 为 $(z_1,\cdots,z_m)\in\mathcal{Z}^m$ 的任意集合。对于上下界为 $[a,a']$ 的有界函数 $\mathcal{G} : \mathcal{Z}\to\mathbb{R}$，假设存在函数 $T : \mathcal{G}\to\mathbb{R}^+$ 及常量 α，因此有任意*

$g \in \mathcal{G}$， $Var(g) \le T(g) \le \alpha \Pr(g)$。假设存在次根函数 ψ、对应的固定点 r^*，对于任意 $r \ge r^*$ 满足

$$\psi(r) \ge \alpha \mathcal{R}(\{g \in \mathcal{G} : T(g) \le r\}). \tag{3.19}$$

对于任意 $K > 1, \delta \in (0,1)$，以至少 $1-\delta$的概率存在如下不等式

$$\Pr(g) \ \le \ \frac{K}{K-1}\widehat{\Pr}(g) + c_1 r^* \ + \ c_2 \frac{\log(1/\delta)}{m}, \tag{3.20}$$

其中，$c_1 = \frac{704K}{\alpha}$， $c_2 = 11(a'-a) + 26K\alpha$， $\Pr(g) = \mathbb{E}[g(z)]$ 为期望值，$\widehat{\Pr}(g) = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{x}_i)$ 为采样样本 $\mathcal{Z}^m$ 上的经验估计。

定理 3.9 (半监督局部 Rademacher 复杂度泛化误差界)**.** 令损失函数满足假设 *3.6*。$\psi(r)$ 为次根函数 *(sub-root function)*，r^* 为 ψ 的固定点。若任意 $r \ge r^*$ 满足

$$\psi(r) \ge \sqrt{2} C_\ell L \mathcal{R}(H_r). \tag{3.21}$$

则对于任意$\delta \in (0,1)$，则以至少 $1-\delta$ 的概率存在

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \le \frac{705}{C_\ell} r^* + \frac{49 C_\ell \log(1/\delta)}{n}, \tag{3.22}$$

其中 $\widehat{f}_n \in H_\kappa$ 为假设空间中经验误差最小的学习器，$f^* \in H_\kappa$ 为假设空间中泛化误差最小的学习器。C_ℓ 为损失函数上界 $\ell \in [0, C_\ell]$。

证明. 根据引理 3.8，对于任意样本$(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$，令 $g = \ell(\widehat{f}_n(\boldsymbol{x}), \boldsymbol{y}) - \ell(f^*(\boldsymbol{x}), \boldsymbol{y})$，其中 $\widehat{f}_n \in H_\kappa$ 为经验损失最小化对应学习器，$f^* \in H_\kappa$ 为期望损失最小化对应学习器. 因此，泛化误差界可以写为 $\Pr(g) = \mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*)$。

第一步: (3.20) → (3.22)。 由于学习器 $\widehat{f}_n$ 对应于最小的经验损失，因此有 $\widehat{\Pr}(g) \le \widehat{\mathcal{E}}(\widehat{f}_n) - \widehat{\mathcal{E}}(f^*) \le 0$。所以可以省略不等式 (3.20) 中的 $\widehat{\Pr}(g)$，可得

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \le \ c_1 r^* + c_2 \frac{\log(1/\delta)}{n}. \tag{3.23}$$

损失函数 ℓ 有界并界定在 $[0, C_\ell]$，引理 3.8 中存在 $g \in [-C_\ell, \ C_\ell]$， $a' = C_\ell, a = -C_\ell$。同时， $\mathcal{E}(f^*)$ 为泛化损失的下确界，可得 $\Pr(g) = \mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \in [0, C_\ell]$。方差满足 $\mathrm{Var}(g) = \Pr(g^2) - [\Pr(g)]^2 \le \Pr(g^2) \le C_\ell \Pr(g)$。令 $T(g) = \Pr(g^2)$，$\alpha = C_\ell$。在不等式 (3.23)中，令 $\alpha = C_\ell$， $a = -C_\ell$， $a' = C_\ell$ 及 $K > 1$ ，可得 (3.22)。

第二步： (3.19) → (3.21)。 将方差设置为 $T(g) = \Pr(g^2)$。因此可得，局部 Rademacher 复杂度 $\mathcal{R}(\{f \in \mathcal{F} : T(g) \leq r\})$ 转变为 $\mathcal{R}(\{\mathbb{E}\,[\ell_{\widehat{f}_n} - \ell_{f^*}]^2 \leq r\})$，其中 $\ell_f = \ell(f(\boldsymbol{x}), \boldsymbol{y})$ 为损失空间中任意损失函数 $\ell \in \mathcal{L}$，样本 $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$，于定义 3.8 中的局部 Rademacher 复杂度定义相同 $\mathcal{R}(H_r)$。同样地，不等式 (3.19) 需要次根函数 ψ_1 以满足

$$\psi_1(r) \geq C_\ell \mathcal{R}(\mathcal{L}_r). \tag{3.24}$$

使用引理 3.7中的收缩熟悉，可得

$$\sqrt{2} C_\ell L \mathcal{R}(H_r) \geq C_\ell \mathcal{R}(\mathcal{L}_r). \tag{3.25}$$

考虑使用满足以下条件的次根函数 $\psi(r)$

$$\psi(r) \geq \sqrt{2} C_\ell L \mathcal{R}(H_r). \tag{3.26}$$

结合 (3.25)、(3.26)，可得次根函数 ψ 满足条件 (3.24)，证得不等式 (3.21). □

传统的由局部 Rademacher 复杂度泛化误差界使用定义在标签相关的损失空间上的 Rademacher 复杂度 $\mathcal{R}(\mathcal{L}_r)$[65]，只能在有标签数据集 D^l 上进行估计，其收敛率通常为 $\mathcal{O}(1/\sqrt{n})$。而定理 3.9 使用定义在标签无关的假设空间上的局部 Rademacher 复杂度 $\mathcal{R}(H_r)$，给出了通用的半监督多输出泛化误差界，可以在有标签数据集 D^l、无标签数据集 D^u 上进行估计，其收敛率通常为 $\mathcal{O}(1/\sqrt{n+u})$。

因此，通过建立有监督的局部 Rademacher 复杂度 $\mathcal{R}(\mathcal{L}_r)$ 和半监督的局部 Rademacher 复杂度 $\mathcal{R}(H_r)$ 的关联，本文将 Rademacher 复杂度的泛化误差分析扩展到半监督学习上，从而使用额外的无标签获得更紧的泛化误差界。

注. 当训练数据中没有无标签数据，即 $u = 0$ 时，局部 Rademacher 复杂度退化为全局 Rademacher复杂度。定理 3.9 中基于局部 Rademacher 复杂度的半监督泛化误差界变为有监督泛化误差界[205]。从定理 3.9 可以看出，泛化误差界取决于 $\mathcal{R}(H_r)$、$\mathcal{O}(1/n)$ 两项，因此泛化误差界的收敛率最快的情况为 $\mathcal{O}(1/n)$。尽管最优收敛率 $\mathcal{O}(1/n)$ 与无标签数据个数 u 无关，但无标签数据可以减小 $\mathcal{O}(1/n)$ 项前的常数，从而使得泛化误差界更紧。

3.3.3 半监督核方法的泛化误差界

本节首先在定理 3.10 中介绍了全部数据（包括有标签数据、无标签数据）上局部 Rademacher 复杂度估计 $\mathcal{R}(H_r)$，该复杂度主要依赖于核函数的尾部特征值之和。然后，通过将局部 Rademacher 复杂度（定理 3.10）带入到通用的泛化误差界（定理 3.9）中，推导出半监督核方法的泛化误差界 (推论 3.11)。

核学习器诱导的特征映射 $\phi : \mathcal{X} \to \mathcal{H}$ 将输入样本从输入空间 $\mathcal{X}$ 映射到再生核希尔伯特空间 $\mathcal{H}$。使用 $\boldsymbol{W}$ 的 $\ell_{2,1}$ 范数来正则化假设空间 $\|\boldsymbol{W}\|_{2,1} = \sum_{k=1}^{K} \|\boldsymbol{W}_{\cdot k}\|_2$，其中 $\boldsymbol{W}_{\cdot k}$ 代表权重矩阵 $\boldsymbol{W}$的第 k 列。

定理 3.10 (半监督核学习器的局部 Rademacher 复杂度)**.** *令核函数由对应积分算子 L_κ 的特征值、特征函数线性表征 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{\infty} \lambda_j \varphi_j(\boldsymbol{x})^\top \varphi_j(\boldsymbol{x})$，其中特征值 $(\lambda_j)_{j=1}^{\infty}$ 呈降序排列。使用 $\|\boldsymbol{W}\|_{2,1} \leq 1$ 作为假设空间正则化。任意 $r > 0$，存在*

$$\mathcal{R}(H_r) \leq 2\sqrt{\frac{1}{n+u} \min_{\theta \geq 0} \left(\frac{\theta r}{4L^2} + \sum_{j > \theta} \lambda_j \right)}. \tag{3.27}$$

证明. 基于收缩引理（引理 3.7），Rademacher 随机变量 ϵ_i 的对称性，及范数不等式 $\|f\|_2 \leq \|f\|_1$，可得

$$\begin{aligned}
&\mathcal{R}(H_r) \\
&= \mathcal{R}(f \in H_\kappa : \mathbb{E}[L^2 \|f - f^*\|_2^2] \leq r) \\
&= \mathcal{R}(f - f^* : f \in H_\kappa, \mathbb{E}[\|f - f^*\|_2^2] \leq \frac{r}{L^2}) \\
&\leq \mathcal{R}(f - g : f, g \in H_\kappa, \mathbb{E}[\|f - g\|_2^2] \leq \frac{r}{L^2}) \\
&= 2\mathcal{R}(f : f \in H_\kappa, \mathbb{E}[\|f\|_2] \leq \frac{\sqrt{r}}{2L}) \\
&\leq 2\mathcal{R}(f : f \in H_\kappa, \mathbb{E}[\|f\|_1] \leq \frac{\sqrt{r}}{2L}).
\end{aligned}$$

令 $\|\boldsymbol{W}\|_p = \|\boldsymbol{W}\|_{2,1} = \sum_{k=1}^{K} \|\boldsymbol{W}_k\|_2$。构造新的假设空间 $H_{2,1}$，以满足如下条件

$$\mathcal{R}(H_r) \leq 2\mathcal{R}(H_{2,1}) \tag{3.28}$$

其中，假设空间 $H_{2,1}$ 构造为

$$H_{2,1} = \{\boldsymbol{x} \to \boldsymbol{W}^\top \phi(\boldsymbol{x}) : \|\boldsymbol{W}\|_{2,1} \leq 1, \mathbb{E}[\|f\|_1] \leq \frac{\sqrt{r}}{2L}\}.$$

对于任意 $\theta \in \mathbb{N}_+^0$，经验 Rademacher 复杂度中右边写为

$$\begin{aligned}
&\frac{1}{n+u}\sum_{i=n+1}^{n+u}\sum_{k=1}^{K}\boldsymbol{\epsilon}_{ik}\langle \boldsymbol{W}_{\cdot k}, \phi(\boldsymbol{x}_i)\rangle \\
=&\frac{1}{n+u}\sum_{k=1}^{K}\left\langle \boldsymbol{W}, \sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{ik}\phi(\boldsymbol{x}_i)\right\rangle \\
=&\sum_{k=1}^{K}\left[\left\langle \sum_{j=1}^{\theta}\sqrt{\lambda_j}\langle \boldsymbol{W}_{\cdot k}, \varphi_j\rangle\varphi_j, \sum_{j=1}^{\theta}\frac{1}{\sqrt{\lambda_j}}\left\langle \frac{1}{n+u}\sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{ik}\phi(\boldsymbol{x}_i), \varphi_j\right\rangle\varphi_j\right\rangle\right. \\
&\left.+\left\langle \boldsymbol{W}_{\cdot k}, \sum_{j>\theta}\left\langle \frac{1}{n+u}\sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{ik}\phi(\boldsymbol{x}_i), \varphi_j\right\rangle\varphi_j\right\rangle\right].
\end{aligned} \tag{3.29}$$

为方便表示，令

$$\Pi_{jk} = \left\langle \frac{1}{n+u}\sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{ik}\phi(\boldsymbol{x}_i), \varphi_j\right\rangle. \tag{3.30}$$

使用等式 (3.30), Cauchy-Schwarz 不等及 Jensen 不等式, 将不等式 (3.29) 写为

$$\begin{aligned}
&\mathcal{R}(H_{2,1}) \\
=&\mathbb{E}\left[\sup_{f\in H_{2,1}}\sum_{i=n+1}^{n+u}\sum_{k=1}^{K}\boldsymbol{\epsilon}_{ik}\left\langle \boldsymbol{W}_{\cdot k}, \phi(\boldsymbol{x}_i)\right\rangle\right] \\
\le&\sup_{f\in H_{2,1}}\sum_{k=1}^{K}\sqrt{\left(\sum_{j=1}^{\theta}\lambda_j\langle \boldsymbol{W}_{\cdot k}, \varphi_j\rangle^2\right)\left(\sum_{j=1}^{\theta}\frac{1}{\lambda_j}\mathbb{E}\left[\Pi_{jk}^2\right]\right)} \\
&+\sum_{k=1}^{K}\|\boldsymbol{W}_{\cdot k}\|_2\sqrt{\sum_{j>\theta}\mathbb{E}\left[\Pi_{jk}^2\right]}.
\end{aligned} \tag{3.31}$$

使用特征值分解，存在 $\mathbb{E}\left[|h_y|\right] = \sqrt{\sum_{j=1}^{\infty}\lambda_j\langle \boldsymbol{W}_{\cdot k}, \varphi_j\rangle^2}$，易得

$$\sum_{k=1}^{K}\sqrt{\sum_{j=1}^{\theta}\lambda_j\langle \boldsymbol{W}_{\cdot k}, \varphi_j\rangle^2} \le \mathbb{E}[\|f\|_1] \le \frac{\sqrt{r}}{2L}. \tag{3.32}$$

将等式 (3.32)、正则化项 $\|\boldsymbol{W}\|_{2,1} \le 1$ 带入等式 (3.31) 中，可得

$$\begin{aligned}
&\mathcal{R}(H_{2,1}) \\
\le&\min_{\theta\ge 0\le n+u}\frac{\sqrt{r}}{2L}\sqrt{\sum_{j=1}^{\theta}\frac{1}{\lambda_j}\mathbb{E}\left[\Pi_{jk}^2\right]}+\sqrt{\sum_{j>\theta}\mathbb{E}\left[\Pi_{jk}^2\right]}.
\end{aligned} \tag{3.33}$$

由特征值分解可得，$\mathbb{E}\,\langle\phi(\boldsymbol{x}),\varphi_j\rangle^2=\lambda_j$。应用 Rademacher 变量的对称性，可得

$$\begin{aligned}&\mathbb{E}\,[\Pi_{jk}^2]\\&=\frac{1}{(n+u)^2}\mathbb{E}\sum_{i,l=1}^{n+u}\boldsymbol{\epsilon}_{ik}\boldsymbol{\epsilon}_{lk}\langle\phi(\boldsymbol{x}_i),\varphi_j\rangle\langle\phi(\boldsymbol{x}_l),\varphi_j\rangle\\&=\frac{1}{(n+u)^2}\sum_{i=n+1}^{n+u}\mathbb{E}\,\langle\phi(\boldsymbol{x}),\varphi_j\rangle^2\\&=\frac{\lambda_j}{n+u}.\end{aligned}\tag{3.34}$$

将不等式 (3.33) 、等式 (3.34) 带入不等式 (3.28)，可得

$$\mathcal{R}(H_r)\ \le\ 2\mathcal{R}(H_{2,1})\ \le\ \min_{\theta>0}\frac{1}{L}\sqrt{\frac{\theta r}{n+u}}+\ 2\sqrt{\sum_{j>\theta}\frac{\lambda_j}{n+u}}.$$

定理得证。 □

定理 3.10 说明期望意义下局部 Rademacher 复杂度 $\mathcal{R}(H_r)$ 由核函数对应积分算子 L_κ 的尾部特征值之和所决定。对应局部 Rademacher 复杂度 $\widehat{\mathcal{R}}(H_r)$ 的经验估计可以使用核矩阵的尾部特征值之和界定。所使用的特征值由截断点 θ 分为两部分：截断点 θ 前的较大特征值、截断点 θ 后的较小特征值。

注. 值得注意的是，由于正则化项 $\boldsymbol{W}\in\mathcal{H}\times\mathbb{R}^K$ (e.g. $\boldsymbol{W}_{2,1}\le 1$) 潜在地与类别个数 K 相关，当类别个数 K 较大时，限制更为严格。

推论 3.11 (半监督核学习器的泛化误差界)**.** 使用核函数 $\sup_{\boldsymbol{x}\in\mathcal{X}}\kappa(\boldsymbol{x},\boldsymbol{x})\le 1$、正则化项 $\|\boldsymbol{W}\|_{2,1}\le 1$，同时令损失函数满足假设 *3.6*。以至少 $1-\delta$ 的概率，存在如下的泛化误差界

$$\mathcal{E}(\widehat{f_n})-\mathcal{E}(f^*)\le\ c_{L,\ell}\left(r^*+\frac{\log(1/\delta)}{n}\right),\tag{3.35}$$

固定点 r^* 存在上界

$$r^*\le\min_{\theta\ge 0}\left(\frac{\theta}{n+u}+\sqrt{\frac{1}{n+u}\sum_{j>\theta}\lambda_j}\right),$$

其中，$\widehat{f_n}\in H_\kappa$ 为假设空间中经验损失最小的学习器，$f^*\in H_\kappa$ 为期望损失最小的学习器。$c_{L,\ell}$ 为仅依赖于 L，C_ℓ 的常数。

上述泛化误差界的最差情况为取全部尾部特征值作，即令 $\theta = 0$，局部 Rademacher 复杂度退化为全局 Rademacher 复杂度，依赖于全部的特征值之和（核函数的迹）。对应泛化误差界 $\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*)$ 的最差情况为

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\left(\sqrt{\frac{1}{n+u}} + \frac{1}{n}\right).$$

有限秩核 当核 κ 存在有限秩 θ，也就是对于下标大于秩的尾部特征值全部为 0，$\forall j > \theta, \lambda_j = 0$。很多核函数满足秩有限的条件，如线性核、多项式核。对于线性核 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$，对应的秩最大为 $\theta = d$。对应多项式核 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + 1)^p$，其对应秩最大为 $\theta = p + 1$。因此，Rademacher 复杂度的收敛率与样本个数而不是样本个数的平方根成反比

$$r^* = \mathcal{O}\left(\frac{\theta}{n+u}\right).$$

指数级衰减特征值 某些核函数的特征值呈指数级下降 $\sum_{j>\theta} \lambda_j = \mathcal{O}(\exp(-\theta))$，如高斯核[44,65]。通过使用 $\theta = \log(n+u)$ 作为截断下标，可得

$$r^* = \mathcal{O}\left(\frac{\log(n+u)}{n+u}\right).$$

秩有限的核、特征值呈指数级衰减的核对应固定点均主要依赖于 $\mathcal{O}(1/(n+u))$，比只使用有标签数据的固定点 $\mathcal{O}(1/n)$ 要紧很多，很大程度上减少了 $\mathcal{O}(1/n)$ 前的常数项。这两种核包括了常见的高斯核、多项式核、线性核，对应的固定点 r^* 数量级较小，在泛化误差界 (3.35) 上不起作用。因此其泛化误差界主要依赖于 (3.35) 中的第二项，而且该项常数较小

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\left(\frac{1}{n}\right).$$

由于局部 Rademacher 复杂度获取了更小的假设空间，通常能够获得更紧的泛化误差界。局部 Rademacher 复杂度的研究已经引起了广泛关注[47,63,214,215]，均使用了基于核矩阵谱分解、核函数谱分解的手段，使用尾部特征值对局部 Rademacher 复杂度进行界定。

3.3.4　半监督线性学习器的泛化误差界

本节首先基于权重矩阵 $\boldsymbol{W}$ 的奇异值分解 (singular values decomposition, SVD)，对线性学习器 $f(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x}$ 的局部 Rademacher 复杂度进行界定。理

论结果 (定理 3.12) 显示局部 Rademacher 复杂度可以使用 $\boldsymbol{W}$ 的尾部奇异值之和进界定。将复杂度估计（定理 3.12）带入到通用泛化误差界（定理 3.9）中，最终得到半监督线性方法的局部 Rademacher 复杂度泛化误差界 (推论 3.13)。

定理 3.12 (半监督线性学习器的局部 Rademacher 复杂度)**.** *对权重矩阵 $\boldsymbol{W}$ 使用奇异值分解，$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$，其中 $\boldsymbol{U}$，$\boldsymbol{V}$ 为特征向量组成的酉阵，$\boldsymbol{\Sigma}$ 为对角矩阵，由降序奇异值 $\{\tilde{\lambda}_j\}$ 构成。令损失函数满足假设 3.6，$\mathbb{E}\,[\boldsymbol{x}^\top\boldsymbol{x}] \le 1$ 及正则化项 $\|\boldsymbol{W}\|_* \le 1$。线性学习器对应的局部 Rademacher 复杂度 $\mathcal{R}(H_r)$ 上界为*

$$\mathcal{R}(H_r) \le 2\sqrt{\frac{1}{n+u}\min_{\theta\ge 0}\Big(\frac{\theta r}{4L^2} + \sum_{j>\theta}\tilde{\lambda}_j^2\Big)}.$$

证明. 基于引理 3.7、Rademacher 变量的对称性以及 $\mathbb{E}\,[\boldsymbol{x}^\top\boldsymbol{x}] \le 1$，可得

$$\begin{aligned}
&\mathcal{R}(H_r)\\
&= \mathcal{R}\big(f \in H_\kappa : \mathbb{E}[L^2\|f - f^*\|_2^2] \le r\big)\\
&= \mathcal{R}\big(f - f^* : f \in H_\kappa, \mathbb{E}[\|f - f^*\|_2^2] \le \frac{r}{L^2}\big)\\
&\le \mathcal{R}\big(f - g : f, g \in H_\kappa, \mathbb{E}[\|f - g\|_2^2] \le \frac{r}{L^2}\big)\\
&= 2\mathcal{R}\big(f : f \in H_\kappa, \mathbb{E}[\|f\|_2^2] \le \frac{r}{4L^2}\big)\\
&= 2\mathcal{R}\big(f : f \in H_\kappa, \mathbb{E}[\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{x}] \le \frac{r}{4L^2}\big)\\
&= 2\mathcal{R}\big(f : f \in H_\kappa, \mathbb{E}[\|\boldsymbol{W}\boldsymbol{W}^\top\|] \le \frac{\sqrt{r}}{2L}\big)\\
&= 2\mathcal{R}\big(H_r^{\boldsymbol{W}}\big).
\end{aligned} \tag{3.36}$$

上述不等式对 $\boldsymbol{W}$ 进行了限制，可得 $\mathbb{E}\,[\|\boldsymbol{W}\boldsymbol{W}^\top\|] \le \frac{\sqrt{r}}{2L}$，可以用于界定 $\boldsymbol{W}$ 相关项。局部 Rademacher 复杂度 $\mathcal{R}\big(H_r^{\boldsymbol{W}}\big)$ 可以写为

$$\begin{aligned}
&\mathcal{R}\big(H_r^{\boldsymbol{W}}\big)\\
&= \mathbb{E}\left[\sup_{f\in H_r^{\boldsymbol{W}}}\frac{1}{n+u}\sum_{i=n+1}^{n+u}\sum_{k=1}^{K}\boldsymbol{\epsilon}_{ik}f_j(\boldsymbol{x}_i)\right]\\
&= \mathbb{E}\left[\sup_{f\in H_r^{\boldsymbol{W}}}\frac{1}{n+u}\sum_{i=n+1}^{n+u}\sum_{k=1}^{K}\boldsymbol{\epsilon}_{ik}\boldsymbol{W}_{\cdot j}^\top\boldsymbol{\phi}(\boldsymbol{x}_i)\right]\\
&= \mathbb{E}\left[\sup_{f\in H_r^{\boldsymbol{W}}}\sum_{k=1}^{K}\boldsymbol{W}_{\cdot j}^\top\left(\frac{1}{n+u}\sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{ik}\boldsymbol{\phi}(\boldsymbol{x}_i)\right)\right]\\
&= \mathbb{E}\left[\sup_{f\in H_r^{\boldsymbol{W}}}\langle\boldsymbol{W}, \boldsymbol{X}_{\boldsymbol{\epsilon}}\rangle\right],
\end{aligned} \tag{3.37}$$

其中 $\boldsymbol{W}, \boldsymbol{X}_{\boldsymbol{\epsilon}} \in \mathbb{R}^{d\times K}$，$\langle \boldsymbol{W}, \boldsymbol{X}_{\boldsymbol{\epsilon}} \rangle = \mathrm{Tr}\big(\boldsymbol{W}^{\top}\boldsymbol{X}_{\boldsymbol{\epsilon}}\big)$ 表示迹范数。将数据相关矩阵 $\boldsymbol{X}_{\boldsymbol{\epsilon}}$ 定义如下:

$$\boldsymbol{X}_{\boldsymbol{\epsilon}} := \left[\frac{1}{n+u}\sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{i1}\phi(\boldsymbol{x}_i), \cdots, \frac{1}{n+u}\sum_{i=n+1}^{n+u}\boldsymbol{\epsilon}_{iK}\phi(\boldsymbol{x}_i)\right].$$

类似于[215] 中定理 5 的证明流程，使用奇异值分解 (SVD) 可得

$$\boldsymbol{W} = \sum_{j\geq 1}\boldsymbol{u}_j\boldsymbol{v}_j^{\top}\tilde{\lambda}_j,$$

其中 $\boldsymbol{u}_j$，$\boldsymbol{v}_j$ 为正交奇异向量 $\boldsymbol{u}_j$、$\boldsymbol{v}_j$。因此，存在如下不等式

$$\begin{aligned}
&\langle \boldsymbol{W}, \boldsymbol{X}_{\boldsymbol{\epsilon}} \rangle \\
\leq& \sum_{j=1}^{\theta}\left\langle \boldsymbol{u}_j\boldsymbol{v}_j^{\top}\tilde{\lambda}_j, \boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\right\rangle + \sum_{j>\theta}\left\langle \boldsymbol{W}, \boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\right\rangle \\
\leq& \left\langle \sum_{j=1}^{\theta}\boldsymbol{u}_j\boldsymbol{v}_j^{\top}\tilde{\lambda}_j, \sum_{j=1}^{\theta}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\right\rangle + \left\langle \boldsymbol{W}, \sum_{j>\theta}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\right\rangle \\
\leq& \left\|\sum_{j=1}^{\theta}\boldsymbol{u}_j\boldsymbol{v}_j^{\top}\tilde{\lambda}_j^2\right\|\left\|\sum_{j=1}^{\theta}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\tilde{\lambda}_j^{-1}\right\| + \|\boldsymbol{W}\|_*\left\|\sum_{j>\theta}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\right\|.
\end{aligned}$$

接下来，对上面不等式中的正则化项进行界定。根据假设空间定义 $H_r^{\boldsymbol{W}}$，存在 $\mathbb{E}[\|\boldsymbol{W}\boldsymbol{W}^{\top}\|] \leq \frac{\sqrt{r}}{2L}$。因此可得

$$\left\|\sum_{j=1}^{\theta}\boldsymbol{u}_j\boldsymbol{v}_j^{\top}\tilde{\lambda}_j^2\right\| = \left\|\sum_{j=1}^{\theta}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\tilde{\lambda}_j^2\right\| \leq \left\|\sum_{j=1}^{\infty}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\tilde{\lambda}_j^2\right\| = \left\|\mathbb{E}[\boldsymbol{W}\boldsymbol{W}^{\top}]\right\| \leq \frac{\sqrt{r}}{2L}. \tag{3.38}$$

使用奇异值分解的性质，易得

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^{\theta}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\tilde{\lambda}_j^{-1}\right\|\right] &= \mathbb{E}\left[\sqrt{\sum_{j=1}^{\theta}\tilde{\lambda}_j^{-2}\langle \boldsymbol{X}_{\boldsymbol{\epsilon}}, \boldsymbol{u}_j\rangle^2}\right] \\
\leq \sqrt{\sum_{j=1}^{\theta}\frac{\tilde{\lambda}_j^{-2}}{n+u}\mathbb{E}[\langle \phi(\boldsymbol{x}), \boldsymbol{u}_j\rangle^2]} &\leq \sqrt{\frac{\theta}{n+u}}.
\end{aligned} \tag{3.39}$$

同样可以得到

$$\mathbb{E}\left[\left\|\sum_{j>\theta}\boldsymbol{X}_{\boldsymbol{\epsilon}}\boldsymbol{u}_j\boldsymbol{u}_j^{\top}\right\|\right] \leq \sqrt{\frac{1}{n+u}\sum_{j>\theta}\tilde{\lambda}_j^2}. \tag{3.40}$$

假设空间 $H_r^{\boldsymbol{W}}$ 中正则化项 $\|\boldsymbol{W}\|_p$ 使用迹范数

$$\|\boldsymbol{W}\|_* \leq 1. \tag{3.41}$$

将之前的不等式 (3.38)，(3.39))，(3.40)，(3.41) 带入到 (3.37) 中，可得

$$\mathcal{R}\big(H_r^{\boldsymbol{W}}\big) = \mathbb{E}\left[\sup_{f\in H_r^{\boldsymbol{W}}}\left\langle \boldsymbol{W}, \boldsymbol{X}_{\boldsymbol{\epsilon}}\right\rangle\right] \le \min_{\theta\ge 0}\left[\frac{1}{2L}\sqrt{\frac{\theta r}{n+u}} + \sqrt{\frac{1}{n+u}\sum_{j>\theta}\tilde{\lambda}_j^2}\right]. \tag{3.42}$$

结合 (3.36)、(3.42)，易得

$$\mathcal{R}(H_r) \le 2\mathcal{R}\big(H_r^{\boldsymbol{W}}\big) \le \min_{\theta\ge 0}\left[\frac{1}{L}\sqrt{\frac{\theta r}{n+u}} + 2\sqrt{\sum_{j>\theta}\frac{\tilde{\lambda}_j^2}{n+u}}\right].$$

定理得证。 □

定理 3.12 估计了多输出的线性学习器的局部 Rademacher 复杂度。从该定理可以看出，不等式右边的第一项 $\theta r/(4L^2)$ 为常量，因此局部 Rademacher 复杂度主要由权重矩阵 $\boldsymbol{W}$ 尾部奇异值平方之和所决定。

注. 对于核化假设空间，局部 Rademacher 复杂度可以由核矩阵 $\mathbf{K}$ 的尾部特征值之和所决定[47,65,205]。类似的，对于线性空间，定理 3.12 说明局部 Rademacher 复杂度可以由线性学习器权重矩阵 $\boldsymbol{W}$ 的尾部奇异值平方之和所决定。

推论 3.13 (半监督线性学习器的泛化误差界)**.** *假设损失函数满足假设 3.6。令 $\mathbb{E}[\boldsymbol{x}^\top\boldsymbol{x}] \le 1$，使用迹范数对假设空间进行正则化 $\|\boldsymbol{W}\|_* \le 1$。对于任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率存在*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \le \tilde{c}_{L,\ell}\left(\tilde{r}^* + \frac{\log(1/\delta)}{n}\right). \tag{3.43}$$

固定点 $\tilde{r}^*$ 由如下不等式所界定

$$\tilde{r}^* \le \min_{\theta\ge 0}\left(\frac{\theta}{n+u} + \sqrt{\frac{1}{n+u}\sum_{j>\theta}\tilde{\lambda}_j^2}\right),$$

其中 $(\tilde{\lambda}_j)_{j=1}^{\infty}$ 为权重矩阵 $\boldsymbol{W}$ 的降序奇异值，$\tilde{c}_{L,\ell}$ 为仅依赖于 L, C_ℓ 的常数。同时，$\widehat{f}_n \in H_\kappa$ 为经验风险最小化 (ERM) 对应的学习器，$f^* \in H_\kappa$ 为期望损失最小化对应的学习器。

局部 Rademacher 复杂度对应的最差情况为考虑全部奇异值 ($\theta = 0$), 退化为全局 Rademacher 复杂度。固定点 $\tilde{r}^*$ 与全局 Rademacher 复杂度相关, 其收敛率通常为 $\mathcal{O}(1/\sqrt{n+u})$。因此，线性学习器泛化误差收敛界为

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\Big(\frac{1}{\sqrt{n+u}} + \frac{1}{n}\Big).$$

与章节 3.3.3 中核学习器对应泛化误差界类似，当 $\boldsymbol{W}$ 对应秩有限、奇异值呈指数级下降时，定理 3.12 的固定点 $\tilde{r}^*$ 主要依赖于 $\tilde{r}^* \leq \theta/(n+u)$。此时，对应的泛化误差界收敛率很快，同时常数较小

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\left(\frac{1}{n}\right).$$

对于线性学习器，Xu 等针对多标签问题提出了类似的理论结果[215]；而已完成工作基于局部 Rademacher 复杂度，推导出线性多分类问题的泛化理论[207]。

注. 权重矩阵的尾部奇异值之和或者核矩阵的尾部特征值之和常用于界定局部 Rademacher 复杂度[47,65,215]。而尾部特征值、尾部奇异值的截断点 θ 对局部 Rademacher 复杂度的估计至关重要，如果截断点过小，局部 Rademacher 复杂度退化为全局 Rademacher 复杂度；如果截断点过大，复杂度估计接近于常数 θ，Cortes 等讨论了不同截断点 θ 对局部 Rademacher 复杂度估计的影响[47]。对于秩有限的核矩阵、权重矩阵 $\boldsymbol{W}$，直接将截断点 θ 设置为秩，此时尾部和恒为 0。对于秩较大的学习器，最理想的截断点 θ 能够让定理 3.10、定理 3.12中不等式右边的两项相等，从而获得最小的固定点。对于核学习器，令 $\frac{\theta r}{4L^2} = \sum_{j>\theta} \lambda_j$；对于线性学习器，令 $\frac{\theta r}{4L^2} = \sum_{j>\theta} \tilde{\lambda}_j^2$。

3.3.5　相关工作比较

本节首先介绍了几种经典的多输出问题 (vector-valued learning) 的数据依赖泛化误差界，并与本文提出的核学习器、线性学习器的泛化误差界进行比较。然后，对多输出问题的两个特例（多分类、多标签）的传统数据依赖泛化误差界进行讨论，并与本文提出的泛化误差界进行比较。

3.3.5.1　一般化的多输出问题的泛化误差界

引理 3.7 中的收缩不等式泛是有监督泛化理论分析的关键，将半监督的局部 Rademacher 复杂代替为半监督的局部 Rademacher 复杂度。

表 3.1 对多输出问题的数据依赖泛化误差界进行了对比，对于核化的多输出学习，Cortes 等[212] 给出的泛化误差收敛率为 $\mathcal{O}\left(\sqrt{\log K/n}\right)$，而 Maurer[213] 给出的泛化误差收敛率为 $\mathcal{O}\left(\sqrt{1/n}\right)$，而定理 3.12、定理 3.10 将泛化误差收敛率提升到至少为 $\mathcal{O}\left(1/\sqrt{n+u}+1/n\right)$。对于线性多输出问题，研究工作[212,213] 的泛化误差收敛率均为 $\mathcal{O}\left(\sqrt{K/n}\right)$，而在推论 3.13 的泛化误差界提供了更快的泛化误差收

表 3.1 多输出问题 (VV) 的泛化误差界对比

泛化误差界结果	最差情况	较好情况
GRC for VV[212]	Kernel: $\mathcal{O}\left(\sqrt{\frac{\log K}{n}}\right)$	Linear: $\mathcal{O}\left(\sqrt{\frac{K}{n}}\right)$
GRC for VV[213]	Kernel: $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$	Linear: $\mathcal{O}\left(\sqrt{\frac{K}{n}}\right)$
LRC for Kernel VV (推论 3.11) †	$\mathcal{O}\left(\frac{1}{\sqrt{n+u}}+\frac{1}{n}\right)$	$\mathcal{O}\left(\frac{1}{n}\right)$
LRC for Linear VV (推论 3.13) †	$\mathcal{O}\left(\frac{1}{\sqrt{n+u}}+\frac{1}{n}\right)$	$\mathcal{O}\left(\frac{1}{n}\right)$

GRC 代表使用了全局 Rademacher 复杂度，而 LRC 代表使用了局部 Rademacher 复杂度。无标签样本个数远大于有标签样本个数，$u \gg n$，而 † 代表泛化误差界使用了无标签数据。

敛率，即使在最差情况也能达到 $\mathcal{O}\left(1/\sqrt{n+u}+1/n\right)$, 该提升主要源于额外无标签数据的使用。另外，在理想情况（如矩阵的秩有限、特征值或奇异值呈指数级下降），核学习器、线性学习器的泛化误差收敛率均为 $\mathcal{O}(1/n)$，比传统的多输出学习的泛化误差界的收敛率[212,213] 要快很多。

相比于传统泛化误差界，本文使用了局部 Rademacher 复杂度、定义在假设空间上标签无关的复杂度估计对传统的泛化误差界 $\mathcal{O}(1/\sqrt{n})$ 进行了提升。使用局部 Rademacher 复杂度代替全局 Rademacher 复杂度，使得在秩有效、特征值或奇异值呈指数级下降的情况，能够获得更快的泛化误差收敛率 $\mathcal{O}(1/n)$。同时，通过使用标签无关的局部 Rademacher 复杂度 $\mathcal{R}(H_r)$ 使用无标签数据获得更小的的 Rademacher 复杂度，从而获得更紧的泛化误差收敛界 $\mathcal{O}(1/\sqrt{n+u})$。

3.3.5.2 特例: 多分类

基于数据依赖的 Rademacher 复杂度，已有众多研究工作对多分类的泛化性能进行分析[216–218]。表 3.2 对多分类的泛化误差界收敛率进行了对比，可以看出，推论 3.11 在核学习器的泛化误差界中，推论 3.13 在线性学习器的泛化误差界中均取得了最快的泛化误差收敛率。

Cortes 等使用全局 Rademacher 复杂度对多分类问题的泛化误差界进行了学习[219]，对应的泛化误差收敛率为 $\mathcal{O}\left(K/\sqrt{n}\right)$。基于高斯复杂度、Slepian 引理，Lei 等将对类别个数 K 的依赖从线性依赖降低到对数依赖[217,220]，对应的泛化误

表 3.2 多分类问题 (MC) 的泛化误差界对比

泛化误差界结果	最差情况	较好情况
GRC for Kernel MC[219]	$\mathcal{O}\left(\frac{K}{\sqrt{n}}\right)$	
GRC for Kernel MC[217,220]	$\mathcal{O}\left(\frac{\log K}{\sqrt{n}}\right)$	
GRC for Kernel MC[218] †	$\mathcal{O}\left(\sqrt{\frac{K}{n}}+K\sqrt{\frac{K}{u}}\right)$	
LRC for Kernel MC[205]	$\mathcal{O}\left(\frac{\log^2 K}{n}\right)$	
LRC for Linear MC[207] †	$\mathcal{O}\left(\frac{K}{\sqrt{n+u}}+\frac{1}{n}\right)$	$\mathcal{O}\left(\frac{1}{n}\right)$
LRC for Kernel VV (推论 3.11) †	$\mathcal{O}\left(\frac{1}{\sqrt{n+u}}+\frac{1}{n}\right)$	$\mathcal{O}\left(\frac{1}{n}\right)$
LRC for Linear VV (推论 3.13) †	$\mathcal{O}\left(\frac{1}{\sqrt{n+u}}+\frac{1}{n}\right)$	$\mathcal{O}\left(\frac{1}{n}\right)$

GRC 代表使用了全局 Rademacher 复杂度，而 LRC 代表使用了局部 Rademacher 复杂度。无标签样本个数远大于有标签样本个数，$u \gg n$，而 † 代表泛化误差界使用了无标签数据。

差收敛率为 $\mathcal{O}(\log K/\sqrt{n})$。 Maximov 等首次提出使用无标签数据对多分类的泛化性能进行提升[218]，收敛率提升至 $\mathcal{O}(\sqrt{K/n}+K\sqrt{K/u})$。

尽管统计泛化理论中广泛使用了定义在整个假设空间上的全局 Rademacher 复杂度，但对学习模型泛化性能起关键作用的往往不是整个函数空间，而且其中假设空间中合适的子空间（通常使用较小方差的函数所构成）[47,65]。因此，局部 Rademacher 复杂度直接定义在最优假设附近的子空间上，对于二分类问题、回归问题均已取得良好的泛化性能[47]。已完成工作首次将局部 Rademacher 复杂度引入到多分类中[205]，首次获得了 $\mathcal{O}(1/n)$ 的泛化误差收敛率；已完成工作[207] 将多分类核学习器的局部 Rademacher 复杂度泛化误差界扩展到线性多分类学习器上，同时使用无标签数据提升泛化性能，得到泛化误差收敛率至少为 $\mathcal{O}(\frac{K}{\sqrt{n+u}}+\frac{1}{n})$；整合之前工作，已完成工作[208] 将多分类泛化误差界拓展到多输出问题上 (vector-valued learning)。

从表 3.2 中可以看出，推论 3.11 在多分类核学习器泛化误差收敛率中取得最优，推论 3.13 在多分类线性学习器泛化误差收敛率中取得最优。

表 3.3 多标签问题 (ML) 的泛化误差界对比

泛化误差界结果	最差情况	较好情况
GRC for Linear ML[221]	$\mathcal{O}(\frac{1}{\sqrt{n}})$	
LRC for Linear ML[215]	$\mathcal{O}(\frac{1}{\sqrt{n}})$	$\mathcal{O}(\frac{1}{n})$
LRC for Kernel VV (推论 3.11) †	$\mathcal{O}(\frac{1}{\sqrt{n+u}}+\frac{1}{n})$	$\mathcal{O}(\frac{1}{n})$
LRC for Linear VV (推论 3.13) †	$\mathcal{O}(\frac{1}{\sqrt{n+u}}+\frac{1}{n})$	$\mathcal{O}(\frac{1}{n})$

GRC 代表使用了全局 Rademacher 复杂度，而 LRC 代表使用了局部 Rademacher 复杂度。无标签样本个数远大于有标签样本个数，$u \gg n$，而 † 代表泛化误差界使用了无标签数据。

3.3.5.3 特例：多标签

表 3.3 对多标签的数据依赖泛化误差界进行了比较，说明通过使用无标签数据，对多标签学习的泛化性能进行了显著提升。

Yu 等首次将 Rademacher 复杂度应用到多标签问题的泛化性能分析中[221]，获得了收敛率为 $\mathcal{O}(1/\sqrt{n})$ 泛化误差界，该工作使用了 $\boldsymbol{W}$ 的迹范数（特征值之和）对全局 Rademacher 复杂度进行了界定；而 Xu 等使用局部 Rademacher 复杂度对多标签的数据依赖泛化误差界进行提升，在矩阵 $\boldsymbol{W}$ 的秩有限、奇异值快速减小时候，收敛率提升至 $\mathcal{O}(1/n)$，文章中使用了尾部奇异值之和对局部 Rademacher 复杂度进行了界定。然而，传统多标签泛化理论仅适用于多标签线性学习器[215,221]，没有给出核学习器的泛化误差界；推论 3.11、推论 3.13 分别给出了核学习器、线性学习器的泛化误差界，同时收敛率优于传统方法。

3.3.6 已完成工作对 Rademacher 复杂度泛化理论研究脉络

已完成工作对基于 Rademacher 复杂度的泛化误差理论已展开深入研究，研究脉络如图 3.1 所示。其中，第一项为 Cortes 等提出的基于全局 Rademacher 复杂度的多分类核方法[219]，其泛化误差收敛率为 $\mathcal{O}(\frac{K}{\sqrt{n}})$；其余三项均为本文已完成的研究成果，包括

(1) 已完成工作[205] 首次将局部 Rademacher 复杂度，并考虑多分类类别间关系，获得了更紧的泛化误差界，收敛率为 $\mathcal{O}(\frac{\log K}{n})$。

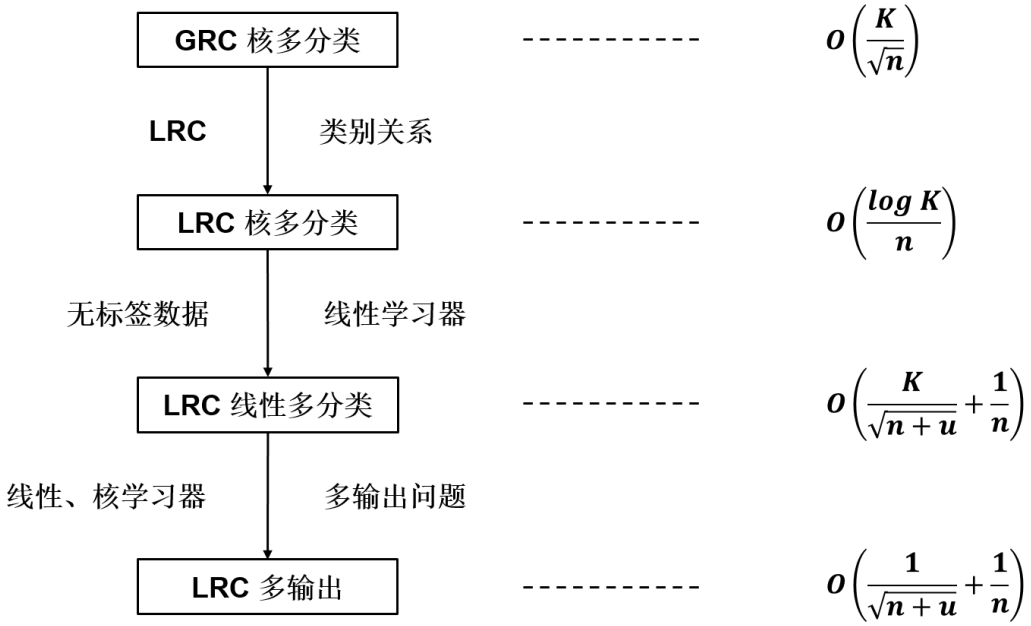

图 3.1 Rademacher复杂度泛化理论研究脉络

(2) 已完成工作[207] 构造半监督 Rademacher 复杂度从而能够使用无标签数据提升多分类泛化性能，并将局部 Rademzher 复杂度理论推广到线性多分类学习器上，得到收敛率为 $\mathcal{O}(\frac{K}{\sqrt{n+u}}+\frac{1}{n})$ 的泛化误差界，u 为无标签数据样本个数。

(3) 已完成工作[107,208] 对之前工作进行整合，之前工作均可视为本工作特例。将多分类局部 Rademacher复杂度泛化理论拓展到多输出问题(vector-valued function)上，并为线性学习器、核学习器提出统一的泛化理论框架，该工作的泛化误差收敛率为 $\mathcal{O}(\frac{1}{\sqrt{n+u}}+\frac{1}{n})$。

3.4　基于积分算子理论的最优泛化理论

核岭回归方法 (kernel ridge regression, KRR) 是一类经典的核方法，常用于拟合非线性回归问题。核岭回归对应单变量标签，即 $\mathcal{Y}=\mathbb{R}$。同时，基于定义 3.1 中假设空间定义 (3.1)，核岭回归 ℓ-2 范数对假设空间进行规范化，即令 $\|\boldsymbol{W}\|_p=\|\boldsymbol{W}\|_{\mathcal{H}}^2$。基于定义 3.4中经验风险最小化的定义，核岭回归使用平方损失 $\ell(f(\boldsymbol{x}),\boldsymbol{y})=(f(\boldsymbol{x})-\boldsymbol{y})^2$ 作为损失函数。核岭回归的经验学习器存在闭式解

$$\widehat{f}_n(\boldsymbol{x})=\sum_{i=1}^{n}\widehat{\alpha}_i\kappa(\boldsymbol{x}_i,\boldsymbol{x}),\quad \text{with}\quad \widehat{\boldsymbol{\alpha}}=(\mathbf{K}+\lambda_A\mathbf{I})^{-1}\mathbf{y}, \tag{3.44}$$

其中$\lambda>0,\mathbf{y}=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)$。核矩阵 $\mathbf{K}$ 为核矩阵，其中元素为 $\mathbf{K}(i,j)=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$。

基于积分算子 (integral operator) 理论，核岭回归的泛化性能已被广泛学习，

并达到最优泛化误差收敛率[69,70]，但由于存储核矩阵需要 $\mathcal{O}(n^2)$ 内存，同时计算等式 (3.44) 中矩阵求逆的时间复杂度为 $\mathcal{O}(n^3)$，所以核岭回归不适用于大规模问题。本节使用积分算子理论中假设空间容量假设（假设 3.14）、假设空间正则化假设（假设 3.15），对两种适用于大规模数据的近似核岭回归算法进行泛化误差分析，并证明结合分布式、随机特征的近似核岭回归仍能达到最优泛化误差收敛率；为结合 Nyström 采样、预处理共轭梯度下降 (PCG) 的半监督核岭回归提供泛化理论保证。

3.4.1 积分算子理论定义与假设

定义 3.10 (积分算子)**.** 核函数 $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 对应积分算子为

$$(L_\kappa g)(\boldsymbol{x}) = \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{z}) d\rho_X(\boldsymbol{z}), \quad \forall g \in L^2(\mathcal{X}, \rho_X),$$

其中 $L^2(\mathcal{X}, \rho_X) = \{h : \mathcal{X} \to \mathbb{R} \mid \|h\|_{\mathcal{H}}^2 = \int |f(\boldsymbol{x})|^2 d\rho_X < \infty\}$，$\rho_X$ 是在输入数据 $\mathcal{X}$ 上的边际分布。

定义 3.11 (有效维 (effective dimension))**.** 给定积分算子 L，有效维定义为

$$\mathcal{N}(\lambda_A) = \operatorname{Tr}\left((L_\kappa + \lambda_A I)^{-1} L_\kappa\right), \quad \lambda_A > 0.$$

由于核函数 κ 是连续、对称、正定的，因此积分算子 L_κ 是紧正迹类算子，同时 $L_\kappa + \lambda_A I$ 可逆。基于积分算子，有效维常用于衡量假设空间 $\mathcal{H}$ 复杂度。

假设 3.14 (容量假设 (capacity assumption))**.** *存在 $Q > 0, \gamma \in [0, 1]$，对于任意 $\lambda_A > 0$，满足*

$$\mathcal{N}(\lambda_A) \leq Q^2 \lambda_A^{-\gamma}.$$

假设 3.15 (正则化假设 (regularity assumption))**.** *存在 $r \in [1/2, 1]$ 及 $g \in L^2(\mathcal{X}, \rho_X)$，满足*

$$f^*(\boldsymbol{x}) = (L_\kappa^r g)(\boldsymbol{x}).$$

其中，f^* 为假设空间中最优模型 $f^* = \arg\min_{f \in H_\kappa} \mathcal{E}(f)$。

假设 3.14 控制了假设空间 $\mathcal{H}$ 的大小，而假设 3.15 对假设空间 f^* 进行规范化。基于积分算子理论，核岭回归的泛化理论分析中广泛使用了假设 3.14、假

设 3.15 这两个假设[68,70,222]，从而获得最优泛化误差率 $\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right)$。Lin 等将积分算子理论拓展到分布式核岭回归的最优泛化误差分析中[75]；Rudi 等将核岭回归最优泛化理论拓展到使用随机特征加速核岭回归的理论分析中[74]。

直观上来说，有效维 $\mathcal{N}(\lambda_A)$ 是估计假设空间 $\mathcal{H}$ 复杂度的度量工具，因此容量假设（假设 3.14）反映了学习器的方差，并等价于经典的熵和覆盖数条件[209]。与 Rademacher 复杂度理论类似，假设 3.14 中 γ 的取值描绘了假设空间 $\mathcal{H}$ 的大小，假设空间越小泛化性能越好。因此 $\gamma = 0$ 对应最优情况，有效维在假设空间的最小子空间上进行度量；$\gamma = 1$ 对应最差情况，有效维在整个假设空间上进行度量。复杂度假设（假设 3.15，也叫 source condition）反映了学习器的偏差，在近似理论中被广泛使用[69]，可以视为对假设空间 f^* 的规范化。假设 3.15 中 r 越大，对假设空间的规范化越严格，对应的学习器越稀疏、泛化性能更好。因此，$\gamma = 1, r = 1/2$ 相当于没有使用假设 3.14、假设 3.15，为最差情况。

3.4.2　结合分布式、随机特征的近似核岭回归泛化分析

3.4.2.1　有监督的泛化误差界

为解决核岭回归的大规模瓶颈问题，本节结合使用分治算法 (divide and conquer)[28]、随机特征 (random feature)[31,74] 加速大规模核岭回归问题的求解。首先将有标签数据集 $D^l = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ 平均划分为 m 块不重叠的数据分块 $\{D_j\}_{j=1}^m$，每个分块上样本个数相同 $|D_1| = \cdots = |D_m| = n/m$。在每个数据分块上使用核岭回归进行求解，最后将所有分块上的解进行平均。同时，在数据分块上使用随机特征 $\kappa(\boldsymbol{x}, \boldsymbol{x}') \approx \phi_M(\boldsymbol{x})^\top \phi_M(\boldsymbol{x}')$ 加速核岭回归的求解。某个数据分块上使用随机特征之后，近似核岭回归的解为

$$\widehat{f}_{D_j}^M(\boldsymbol{x}) = \phi_M(\boldsymbol{x})^\top \boldsymbol{W}_j, \quad \text{with} \quad \boldsymbol{W}_j = (\widehat{S}_M^\top \widehat{S}_M + \lambda_A I)^{-1} \widehat{S}_M^\top \mathbf{y}, \tag{3.45}$$

其中，对于第 j 个数据分块 D_j，$\forall (\boldsymbol{x}, y) \in D_j, \widehat{S}_M^\top = \frac{1}{\sqrt{n/m}}(\phi_M(\boldsymbol{x}_1), \cdots, \phi_M(\boldsymbol{x}_{n/m}))$、$\widehat{\mathbf{y}} = \frac{1}{\sqrt{n/m}}(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{n/m})$。将每个分块上求得的解进行平均，得到最终解

$$\widehat{f}_n = \frac{1}{m} \sum_{j=1}^m \widehat{f}_{D_j}^M. \tag{3.46}$$

定理 3.16 (结合分布式、随机特征的近似核岭回归泛化误差界)**.** *假定结合分布式、随机特征的近似核岭回归方法 (KRR-DC-RF) 满足假设 3.14、假设 3.15，同时*

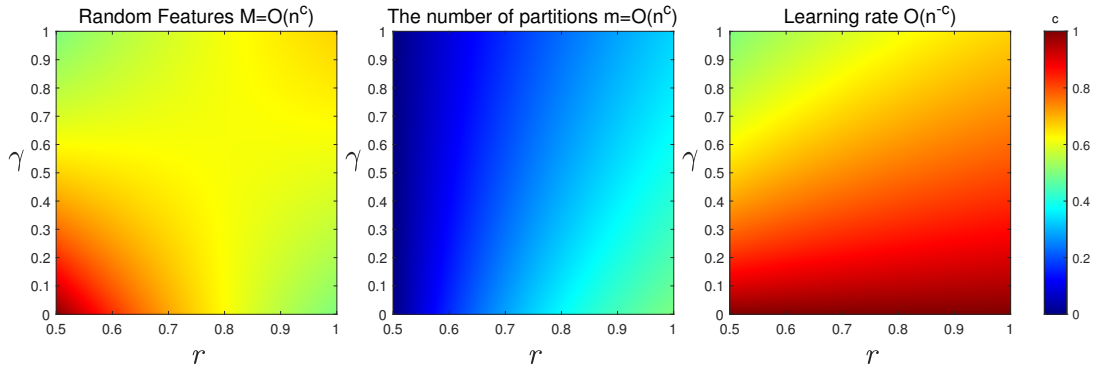

图 3.2 不同假设对随机特征个数 M、分块数 m、泛化误差收敛率的影响

令正则化系数为 $\lambda_A = n^{-\frac{1}{2r+\gamma}}$，同时随机特征个数 M、分块数 m 分别满足

$$M \gtrsim n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \quad m \lesssim n^{\frac{2r-1}{2r+\gamma}},$$

则以很大概率存在如下的泛化误差收敛界

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

其中 $\widehat{f}_n \in H_\kappa$ 为假设空间中经验误差最小的学习器，$f^* \in H_\kappa$ 为假设空间中泛化误差最小的学习器。

通过使用容量假设（假设 3.14）、正则化假设（假设 3.15），Caponnetto 和 De Vito[70] 证明了原始核岭回归方法存在泛化误差上界收敛率为 $\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right)$，而研究工作[70,223] 证明了泛化误差下界的收敛率也为 $\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right)$，因此该泛化误差收敛率是极小极大最优的 (minimax optimal)。近期研究证明了核岭回归的近似方法也能达到最优泛化误差收敛率：(1) 结合分布式的核岭回归方法 (KRR-DC) 在分块数满足 $m \lesssim n^{\frac{2r-1}{2r+\gamma}}$ 时能够获得最优泛化误差收敛率[75]；(2) 结合随机特征的核岭回归方法 (KRR-RF) 在随机特征数满足 $M \gtrsim n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}$ 时能够获得最优泛化收敛率[74]；(3) Lin 和 Rosasco 证明在步长、迭代次数、批次大小满足一定情况下，使用随机梯度方法求解的核岭回归方法仍能达到最优泛化误差收敛率[224]。

学习任务本身的难度、学习算法的匹配程度决定了假设 3.14、假设 3.15 中 γ, r 的取值，对应于不同的泛化误差性能，其泛化误差收敛率在 $\mathcal{O}(1/\sqrt{n})$ 与 $\mathcal{O}(1/n)$ 之间。图 3.2 显示了假设 3.14、假设 3.15 中不同取值的 r、γ 对分块数 m、随机特征个数 M 的影响，每个子图的右下方向代表了更紧的正则化、更小的假设空间，对应于更快的收敛率。在最好情况中，$r = 1, \gamma = 0$（更紧的正则

化、更小假设空间），$\mathcal{O}(\sqrt{n})$ 的随机特征、$\mathcal{O}(\sqrt{n})$ 分块数，就能够达到 $\mathcal{O}(1/n)$ 泛化误差收敛率；而最差情况中，$r=0,\gamma=1$ （不使用容量假设、正则化假设），$\mathcal{O}(\sqrt{n})$ 的随机特征、$\mathcal{O}(1)$ 分块数，就能够达到 $\mathcal{O}(1/\sqrt{n})$ 泛化误差收敛率。

注. 图 3.2 的中间子图显示，不使用假设 3.14、假设 3.15 (即 $r=1/2,\gamma=1$) 对应于最差情况，此时 $\mathcal{O}(\sqrt{n})$ 随机特征、常数级的分块数能够得到较差的泛化误差收敛率 $\mathcal{O}(1/\sqrt{n})$。分块数为常数级 $m=\mathcal{O}(1)$，极大地限制了分布式算法的应用。在接下来的定理 3.17 使用无标签数据对分块数进行提升。

3.4.2.2　半监督的泛化误差界

在引理 3.19 的误差分解显示额外的无标签数据能够显著地减少经验损失 (empirical error)、分块损失 (distributed error)，从而放松对分块数 m 的限制。下面介绍 Chang 等提出的半监督核岭回归框架[225]：将全部数据 $D^l\cup D^u$ 均分为 m 份，对应第 j 块数据为

$$D_j^*=\{D_j^l\cup D_j^u\}_{j=1}^m \quad \text{with}$$

$$\boldsymbol{x}_i^*=\begin{cases}\boldsymbol{x}_i, & \text{if}(\boldsymbol{x}_i,\boldsymbol{y}_i)\in D_j^l,\\ \widetilde{\boldsymbol{x}}_i, & \text{其他},\end{cases}\quad\text{and}\quad \boldsymbol{y}_i^*=\begin{cases}\frac{|D_j^*|}{|D_j|}\boldsymbol{y}_i, & \text{if}(\boldsymbol{x}_i,\boldsymbol{y}_i)\in D_j^u,\\ 0, & \text{其他}.\end{cases}$$

令 $D^*=\bigcup_{j=1}^m D_j^*$, $|D^*|=(n+u)$, $|D_1^*|=\cdots=|D_m^*|=(n+u)/m$。结合分布式、随机特征的半监督核岭回归方法 (SKRR-DC-RF) 定义为

$$\widehat{f}_n=\frac{1}{m}\sum_{j=1}^m \widehat{f}_{D_j^*}^M. \tag{3.47}$$

定理 3.17. 假定 *SKRR-DC-RF* 满足假设 *3.14*、假设 *3.15*、$\lambda_A=n^{-\frac{1}{2r+\gamma}}$, 同时随机特征个数 M，分块数 m 满足

$$M\gtrsim n^{\frac{(2r-1)\gamma+1}{2r+\gamma}},\qquad m\lesssim\min\left\{n^{\frac{2r+2\gamma-1}{2r+\gamma}},\,(n+u)n^{\frac{-\gamma-1}{2r+\gamma}}\right\}.$$

则以较高概率存在如下的泛化误差界，

$$\mathcal{E}(\widehat{f}_n)-\mathcal{E}(f^*)=\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right).$$

其中 $\widehat{f}_n\in H_\kappa$ 为假设空间中经验误差最小的学习器，$f^*\in H_\kappa$ 为假设空间中泛化误差最小的学习器。

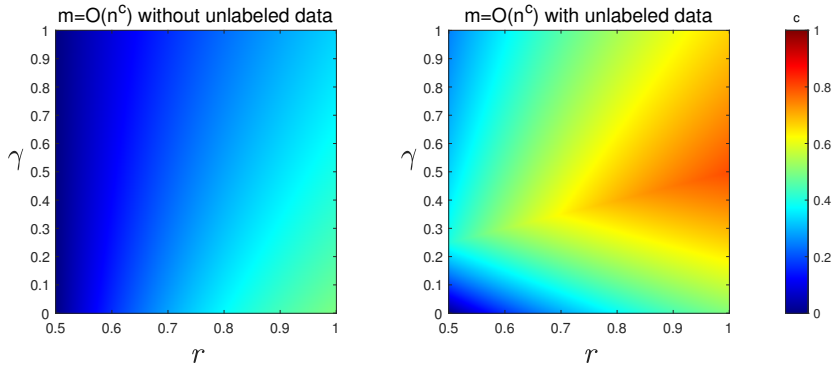

图 3.3 无标签数据对分块数、泛化误差收敛率的影响

当不使用无标签数据，即 $u = 0$，定理 3.17 与定理 3.16 结果一致。由于已经达到了最优泛化误差收敛率，额外无标签数据的引入不会对泛化误差界有提升，但可以用于降低对分块数 m 的依赖。考虑 $(n+u) = n^{1+\frac{r}{2r+\gamma}}$ 的情况，此时 $(n+u) \in [n^{1.25}, n^{1.5}]$ 是大规模半监督回归数据的常见情况。

在图 3.3 中，令总样本数为 $(n+u) = n^{1+\frac{r}{2r+\gamma}}$，图中讨论了分块数 m、总样本数 $(n+u)$ 随着 r, γ 的变化情况。左子图仅使用有标签数据；右子图使用额外的无标签数据。对比图 3.3 两个子图，说明使用无标签数据之后分块数 m 增长很多；同时，无标签数据的使用有效的避免了常数分块数，此时只有 $r = 1/2$、$\gamma = 0$ 一种情况能使得分块数 m 为常数 $\mathcal{O}(1)$。

推论 3.18 (使用无标签数据之后的最差情况)**.** 假定 $y \le |b|, b > 0, n \ge n_0, \lambda_A = n^{-1/2}$，同时随机特征数 M，分块数 m 满足

$$M \gtrsim \sqrt{n}, \quad m \lesssim \min\left\{n, \frac{(n+u)}{n}\right\}.$$

则以较高概率取得

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

其中 $\widehat{f}_n \in H_\kappa$ 为假设空间中经验误差最小的学习器，$f^* \in H_\kappa$ 为假设空间中泛化误差最小的学习器。

推论 3.18 中的泛化误差收敛率仅为 $\mathcal{O}(1/\sqrt{n})$，对应于标准核岭回归的最差情况。不妨设总样本个数为 $n + u = n^{1+\beta}$, $\beta > 0$，则分块数为 $m = \mathcal{O}(n^\beta)$、随机特征数为 $\mathcal{O}(\sqrt{n})$，能够达到泛化误差收敛率 $\mathcal{O}(1/\sqrt{n})$。将分块数 m 与无标签样本数通过 n^β 关联起来，随着额外无标签样本个数的增加，分块数急剧增加。在没有泛化性能损失的情况下，无标签数据的使用显著地提高了计算效率。

3.4.2.3　证明思路

首先将 (3.47) 中结合分治、随机特征的半监督核岭回归学习器 (SKRR-DC-RF) $\widehat{f}_{D_j^*}^M$ 写为原问题形式，同时引入其他几个学习器作为泛化分析的工具：

$$\widehat{f}_n = \frac{1}{m}\sum_{j=1}^{m}\langle \widehat{w}_j, \phi_M(\cdot)\rangle,\ \widehat{w}_j = \arg\min_{w\in\mathbb{R}^M}\left\{\frac{1}{n+u}\sum_{i=1}^{n+u}(\langle w, \phi_M(\boldsymbol{x}_i^*)\rangle - y_i^*)^2 + \lambda_A\|w\|^2\right\},$$

$$\widetilde{f}_{D^*}^M = \frac{1}{m}\sum_{j=1}^{m}\langle \widetilde{w}_j, \phi_M(\cdot)\rangle,\ \widetilde{w}_j = \arg\min_{w\in\mathbb{R}^M}\left\{\frac{1}{n+u}\sum_{i=1}^{n+u}\left(\langle w, \phi_M(\boldsymbol{x}_i^*)\rangle - f^*(\boldsymbol{x}_i^*)\right)^2 + \lambda_A\|w\|^2\right\},$$

$$f_{\lambda_A}^M = \langle \widehat{u}, \phi_M(\cdot)\rangle,\ u = \arg\min_{u\in\mathbb{R}^M}\int_{\mathcal{X}}\left(\langle u, \phi_M(\boldsymbol{x})\rangle - f^*(\boldsymbol{x})\right)^2 d\rho_X(\boldsymbol{x}) + \lambda_A\|u\|^2,$$

$$f_{\lambda_A} = \langle \widehat{u}, \phi(\cdot)\rangle,\ v = \arg\min_{v\in\mathcal{H}_K}\int_{\mathcal{X}}\left(\langle v, \phi(\boldsymbol{x})\rangle - f^*(\boldsymbol{x})\right)^2 d\rho_X(\boldsymbol{x}) + \lambda_A\|v\|^2,$$

其中$\phi : \mathbb{R}^d \to \mathcal{H}$ 为核函数诱导的隐式特征映射 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}')\rangle$。经验学习器 $\widetilde{f}_{D^*}^M$ 在无噪声的数据上进行学习；期望学习器 $f_{\lambda_A}^M$ 使用了随机特征 ϕ_M 对核学习器进行近似；而期望学习器 f_{λ_A} 使用了核函数诱导的隐式随机特征 ϕ。研究工作[70,226] 中给出了如下不等式，将泛化误差界与以学习器差异的范数联系起来

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) = \mathbb{E}\,\|\widehat{f}_n - f^*\|_{\mathcal{H}}^2. \tag{3.48}$$

经验误差最小化学习器与期望误差最小化学习器的差异可以划分为

$$\widehat{f}_n - f^* = \widehat{f}_n - f_{\lambda_A}^M + f_{\lambda_A}^M - f_{\lambda_A} + f_{\lambda_A} - f^*,$$

结合不等式 (3.48)，可以推导出如下引理

引理 3.19. *基于 $\widehat{f}_n, \widetilde{f}_{D^*}^M, f_{\lambda_A}^M$，$f_{\lambda_A}$ 等学习器定义，存在如下泛化误差界分解为*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \tag{3.49}$$

$$\le \frac{6}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}_{D_j^*}^M - \widetilde{f}_{D_j^*}^M\|_{\mathcal{H}}^2 \quad (\textit{Variance}) \tag{3.50}$$

$$+ \frac{6}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j^*}^M - f_{\lambda_A}^M\|_{\mathcal{H}}^2 \quad (\textit{Empirical error}) \tag{3.51}$$

$$+ \frac{3}{m}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j^*}^M - f_{\lambda_A}^M\|_{\mathcal{H}}^2 \quad (\textit{Distributed Error}) \tag{3.52}$$

$$+\ 3\,\|f_{\lambda_A}^M - f_{\lambda_A}\|_{\mathcal{H}}^2 \quad (\textit{Random Features Error}) \tag{3.53}$$

$$+\ 3\,\|f_{\lambda_A} - f^*\|_{\mathcal{H}}^2 \quad (\textit{Approximation Error}). \tag{3.54}$$

证明. 将划分

$$\widehat{f_n} - f^* = \widehat{f_n} - f_{\lambda_A}^M + f_{\lambda_A}^M - f_{\lambda_A} + f_{\lambda_A} - f^*,$$

引入到等式 (3.48) 中，并由 $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$，可得

$$\mathbb{E}\big[\mathcal{E}(\widehat{f_n}) - \mathcal{E}(f^*)\big] \le 3\,\mathbb{E}\|\widehat{f_n} - f_{\lambda_A}^M\|_{\mathcal{H}}^2 + 3\,\mathbb{E}\|f_{\lambda_A}^M - f_{\lambda_A}\|_{\mathcal{H}}^2 + 3\,\mathbb{E}\|f_{\lambda_A} - f^*\|_{\mathcal{H}}^2. \quad (3.55)$$

下面对 $\mathbb{E}\|\widehat{f_n} - f_{\lambda_A}^M\|_{\mathcal{H}}^2$ 进一步分解

$$\begin{aligned}
&\|\widehat{f_n} - f_{\lambda_A}^M\|_{\mathcal{H}}^2 \\
&= \left\|\frac{1}{m}\sum_{j=1}^{m}(\widehat{f}_{D_j^*}^M - f_{\lambda_A}^M)\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{m^2}\sum_{j=1}^{m}\|(\widehat{f}_{D_j^*}^M - f_{\lambda_A}^M)\|_{\mathcal{H}}^2 + \frac{1}{m}\sum_{j=1}^{m}\left\langle \widehat{f}_{D_j^*}^M - f_{\lambda_A}^M, \frac{1}{m}\sum_{k\neq j}(\widehat{f}_{D_k,\lambda_A}^M - f_{\lambda_A}^M)\right\rangle_{\mathcal{H}} \\
&= \frac{1}{m^2}\sum_{j=1}^{m}\|(\widehat{f}_{D_j^*}^M - f_{\lambda_A}^M)\|_{\mathcal{H}}^2 + \frac{1}{m}\sum_{j=1}^{m}\left\langle \widehat{f}_{D_j^*}^M - f_{\lambda_A}^M, \widehat{f_n} - f_{\lambda_A}^M - \frac{1}{m}(\widehat{f}_{D_j^*}^M - f_{\lambda_A}^M)\right\rangle_{\mathcal{H}}.
\end{aligned}$$

对 $\|\widehat{f_n} - f_{\lambda_A}^M\|_{\mathcal{H}}^2$ 求期望可得

$$\begin{aligned}
&\mathbb{E}\|\widehat{f_n} - f_{\lambda_A}^M\|_{\mathcal{H}}^2 \\
&= \frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\|(\widehat{f}_{D_j^*}^M - f_{\lambda_A}^M)\|_{\mathcal{H}}^2 \\
&+ \frac{1}{m}\sum_{j=1}^{m}\left\langle \mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M, \mathbb{E}[\widehat{f_n}] - f_{\lambda_A}^M - \frac{1}{m}(\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M)\right\rangle_{\mathcal{H}}.
\end{aligned}$$

上式右边的第二部分可以写为

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\left\langle \mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M, \mathbb{E}[\widehat{f_n}] - f_{\lambda_A}^M\right\rangle_{\mathcal{H}} \\
&- \frac{1}{m}\sum_{j=1}^{m}\left\langle \mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M, \frac{1}{m}(\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M)\right\rangle_{\mathcal{H}} \\
&= \|\mathbb{E}[\widehat{f_n}] - f_{\lambda_A}^M\|_{\mathcal{H}}^2 - \frac{1}{m^2}\sum_{j=1}^{m}\|\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M\|_{\mathcal{H}} \\
&= \left\|\frac{1}{m}\sum_{j=1}^{m}(\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M)\right\|_{\mathcal{H}}^2 - \frac{1}{m^2}\sum_{j=1}^{m}\|\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M\|_{\mathcal{H}}.
\end{aligned}$$

使用 Cauchy–Schwarz 不等式，可得

$$\left\|\frac{1}{m}\sum_{j=1}^{m}(\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M)\right\|_{\mathcal{H}}^2 \le \frac{1}{m}\sum_{j=1}^{m}\|(\mathbb{E}[\widehat{f}_{D_j^*}^M] - f_{\lambda_A}^M)\|_{\mathcal{H}}^2. \quad (3.56)$$

使用 Jensen's 不等式，可得

$$\frac{1}{m}\sum_{j=1}^{m}\|(\mathbb{E}[\widehat{f}_{D_j^*}^M]-f_{\lambda_A}^M)\|_{\mathcal{H}}^2 \le \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|(\widetilde{f}_{D_j}^M-f_{\lambda_A}^M)\|_{\mathcal{H}}^2.$$

最终，结合不等式 (3.56) 的第一部分，易得

$$\mathbb{E}\|\widehat{f}_n-f_{\lambda_A}^M\|_{\mathcal{H}}^2 \le \underbrace{\frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}_{D_j^*}^M-f_{\lambda_A}^M\|_{\mathcal{H}}^2}_{\text{Sample Error}} + \underbrace{\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j}^M-f_{\lambda_A}^M\|_{\mathcal{H}}^2}_{\text{Distributed Error}}.$$

泛化误差可以划分为采样误差 (sample error)、分布式学习器带来的误差 (distributed error)。将 $\|\widehat{f}_{D_j^*}^M-f_{\lambda_A}^M\|_{\mathcal{H}}^2$ 划分为 $\|\widehat{f}_{D_j^*}^M-\widetilde{f}_{D_j}^M+\widetilde{f}_{D_j}^M-f_{\lambda_A}^M\|_{\mathcal{H}}^2$，并根据不等式 $(a+b)^2 \le 2a^2+2b^2$ 可得

$$\begin{aligned}&\mathbb{E}\,\|\widehat{f}_n-f_{\lambda_A}^M\|_{\mathcal{H}}^2 \le \\ &\underbrace{\frac{2}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widehat{f}_{D_j^*}^M-\widetilde{f}_{D_j}^M\|_{\mathcal{H}}^2}_{\text{Variance}} + \underbrace{\frac{2}{m^2}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j}^M-f_{\lambda_A}^M\|_{\mathcal{H}}^2}_{\text{Empirical error}} + \underbrace{\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|\widetilde{f}_{D_j}^M-f_{\lambda_A}^M\|_{\mathcal{H}}^2}_{\text{Distributed Error}}.\end{aligned} \tag{3.57}$$

值得注意的是，$\widehat{f}_{D^*,\lambda_A}^M-f_{\lambda_A}^M$ 的范数包含了模型方差 (variance)，数据集采样带来的经验误差 (empirical error)，以及分布式学习器带来的误差 (distributed error)。将不等式 (3.57) 带入 (3.55) 中，可以证得引理。 □

引理 3.19 中将泛化误差划分为模型方差（来源于样本噪声）、经验误差（来源于样本采样）、分布式误差（来源于分治算法近似）、随机特征误差（来源于随机特征近似核函数）、近似误差（来源于学习器与模型最优学习器的偏差）。之后，使用积分算子理论手段对对误差分解得到的各项进行界定，从而证得定理 3.17，而定理 3.16 是在其仅使用有标签数据的特例 $u=0$。详细证明过程在已完成工作[188] 中的定理 25 、定理 26 的证明中给出。

3.4.3　结合 Nyström 采样、PCG的近似 LapRLS 泛化分析

考虑使用基于流形假设的半监督核岭回归，对应于拉普拉斯最小二乘法 (Laplacian regularized least squares，LapRLS)。 LapRLS 使用平方损失作为损失函数，求解如下的优化目标

$$\arg\min_{f\in\mathcal{H}}\sum_{i=1}^{n}(\boldsymbol{y}_i-f(\boldsymbol{x}_i))^2+\lambda_A\|f\|_{\mathcal{H}}^2+\lambda_I\mathbf{f}^{\top}\mathbf{L}\mathbf{f}. \tag{3.58}$$

其中，图 Laplacian 矩阵定义为 $\mathbf{L}=\mathbf{D}-\boldsymbol{S}$。 $\boldsymbol{S}$ 为相似度矩阵，$\boldsymbol{S}_{ij}$ 衡量了样本输入 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 之间的相似性。通常由 k-近邻方法选取 k 个最近邻的数据输入计算相似性，而其他元素设置为 0。 $\mathbf{D}$ 为对角矩阵，对角元素为 $\mathbf{D}_{ii}=\sum_{j=1}^{n+u}\boldsymbol{S}_{ij}$。同时有 $\mathbf{f}=[f(\boldsymbol{x}_1),\cdots,f(\boldsymbol{x}_n)]^\top$，$\lambda_A$。基于表示定理[227]

$$f(\boldsymbol{x})=\sum_{i=1}^{n}\alpha_i\kappa(\boldsymbol{x}_i,\boldsymbol{x}).$$

将优化目标的梯度设置为 0，可得 LapRLS 的闭式解

$$\widehat{\boldsymbol{\alpha}}=(\mathbf{JK}+\lambda_A\mathbf{I}+\lambda_I\mathbf{LK})^{-1}\mathbf{y}, \tag{3.59}$$

其中$\mathbf{K}_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 为 $(n+u)\times(n+u)$ 大小的核矩阵。$\mathbf{J}=\mathrm{diag}(1,\cdots,1,0,\cdots,0)$，其中前 n 个对角线元素为 1 其他为 0，$\mathbf{y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n,0,\cdots,0]^\top$ 前 n 个元素使用 D^l 中标签，而剩余 u 个元素使用 0 填充。 λ_A 为假设空间正则化系数，λ_I 为 Laplacian 正则化系数。将 Laplacian 正则化系数设置为 $\lambda_I=0$，对应 LapRLS 解 (3.59) 中不使用无标签数据，退化为标准核岭回归。

3.4.3.1 Nyström–RLS 方法的泛化误差界

LapRLS 同样存在大规模瓶颈，使用 Nyström 采样来减少内存需求、提高计算效率，Nyström 方法使用采样后的较小核矩阵对原始核矩阵进行近似。使用 Nyström 方法采样获得 s 个样本，对应于某个假设空间子空间

$$H_s=\{f\in H_\kappa|f=\sum_{i=1}^{s}\alpha_i\kappa(\boldsymbol{x}_i,\cdot),\boldsymbol{\alpha}\in\mathbb{R}^s\},$$

其中$s\le n+u$。 $\boldsymbol{x}_s=(\widetilde{\boldsymbol{x}}_1,\cdots,\widetilde{\boldsymbol{x}}_s)$ 为训练数据集上通过 Nyström 采样获得的数据点。使用 Nyström 采样后，从 (3.58) 推导得出 Nyström–LapRLS 的解为:

$$\widehat{f_n}=\sum_{i=1}^{s}\alpha_i\kappa(\boldsymbol{x}_i,\boldsymbol{x}),\qquad \boldsymbol{\alpha}=\underbrace{(\mathbf{K}_{ns}^\top\mathbf{K}_{ns}+\lambda_A\mathbf{K}_{ss}+\lambda_I\mathbf{K}_{\cdot s}^\top\mathbf{LK}_{\cdot s})^\dagger}_{\mathbf{H}}\underbrace{\mathbf{K}_{ns}^\top\mathbf{y}}_{\mathbf{z}}, \tag{3.60}$$

其中$\mathbf{H}^\dagger$ 为矩阵 $\mathbf{H}$ 的 Moore-Penorse 伪逆。全部数据上的核矩阵为 $[\mathbf{K}]_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$，其中 $\forall\, i,j\in\{1,\cdots,n+u\}$。定义在有标签数据、采样数据上的核矩阵 $\mathbf{K}_{ns}$ 为 $\mathbf{K}$ 的一部分，行由 n 个有标签数据所对应下标确定，列由 s 个采样数据对应下标确定。定义在采样数据上的核矩阵 $\mathbf{K}_{ss}$ 为 $\mathbf{K}$ 的一部分，行由 s 个采样数据所对应下标确定，列由 s 个采样数据对应下标确定。 $\mathbf{K}_{\cdot s}$ 为

为 $\mathbf{K}$ 的一部分，行为全部行，列由 s 个采样数据对应下标确定。标签向量为 $\mathbf{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]^\top \in \mathbb{R}^n$。

为 Nyström–RLS 引入通用的多正则化乡学习框架

$$\widehat{f_n} = \arg\min_{f \in H_s} \frac{1}{n} \sum_{i=1}^{n} ||f(\boldsymbol{x}_i) - \boldsymbol{y}_i||_{\mathcal{Y}}^2 + \lambda_A ||f||_{\mathcal{H}}^2 + \sum_{j=1}^{p} \beta_j ||B_j f||_{\mathcal{H}}^2,$$

其中$B_j : \mathcal{H} \to \mathcal{H}$ $(1 \le j \le p)$ 为有界算子，$\lambda_A > 0$，β_j $(1 \le j \le p)$ 为非负实数。考虑最差情况 $(\gamma = 0, r = 1/2)$，即不使用假设 3.14、假设 3.15。已有工作[135] 的定理 2中给出如下的 Nyström–RLS 泛化误差界。

定理 3.20 (Nyström–LapRLS 的泛化误差界)**.** *存在常量 M, σ，令 $\lambda_A \ge \frac{8C_\kappa^2}{\sqrt{n}} \log\left(\frac{4}{\delta}\right)$，并使用足够多的 Nyström 采样样本 $s \ge \max\left\{67 \log\left(\frac{12C_\kappa^2}{\lambda_A \delta}\right), 2C_\kappa^2 \log\left(\frac{12C_\kappa^2}{\lambda_A \delta}\right)\right\}$。任意 $\delta \in (0,1)$，以至少 $1-\delta$ 取得如下泛化误差界*

$$\left[\mathcal{E}(\widehat{f_n}) - \mathcal{E}(f^*)\right]^{1/2} \le c_3 {\lambda_A}^r + c_4 \frac{\mathcal{B}_{\lambda_A}}{\lambda_A} + \left\{\frac{8C_\kappa M}{n\sqrt{\lambda_A}} + 8\sigma\sqrt{\frac{\mathcal{N}(\lambda_A)}{n}}\right\} \log\left(\frac{6}{\delta}\right), \quad (3.61)$$

其中常量 c_3, c_4 为常量，$\mathcal{B}_{\lambda_A} = ||\sum_{j=1}^{p} \beta_j B_j^* B_j||$。

3.4.3.2　结合 PCG 的Nyström–RLS 方法的泛化误差界

Nyström–LapRLS 的解 (3.60) 为线性方法，包括了伪逆部分，而直接求解伪逆 $\mathbf{H}^\dagger$ 时间复杂度较高。考虑使用预处理共轭梯度 (preconditioned conjugate gradient，PCG) 算法对线性系统的求解进行加速

$$\mathbf{P}^{-1}\mathbf{H}\boldsymbol{\alpha} = \mathbf{P}^{-1}\mathbf{z}.$$

其中，$\mathbf{P}$ 为预处理器。使用迭代算法求解线性系统的迭代次数取决于条件数 (condition number)，条件数定义为 $cond(\mathbf{P}^{-1}\mathbf{H}) = \frac{\sigma_{\max}(\mathbf{H})}{\sigma_{\min}(\mathbf{H})}$。较小的条件数对应良置的求解问题，迭代次数会更少，因此预处理器 $\mathbf{P}$ 应与 $\mathbf{H}$ 尽量相近。将经验学习器 $\widehat{f_n}$ 定义为使用 PCG 方法求解 (3.60)。

定理 3.21. *令 $n_0 \in \mathbb{N}$，假设 $|y| \le b$, $\forall b > 0$，$n \ge n_0$*

$$\lambda_A = \frac{8C_\kappa^2}{\sqrt{n}} \log\left(\frac{4}{\delta}\right), \quad s \ge 5(67 + 20\sqrt{n}) \log\frac{48C_\kappa^2 n}{\delta}$$
$$t \ \ge \ \frac{1}{2}\log n + 2\log(2b + 3C_\kappa) + 5.$$

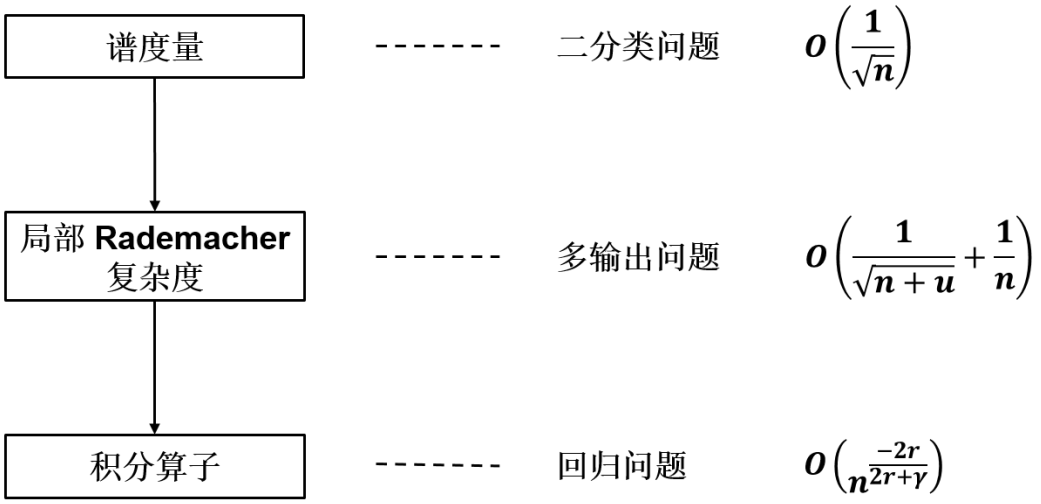

图 3.4 大规模半监督核方法泛化理论研究脉络

任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率取得如下泛化误差界

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \leq \frac{c_5 \ \log^2 \frac{24}{\delta}}{\sqrt{n}},$$

其中 n_0, c_5 为常量。$\widehat{f}_n \in H_\kappa$ 为假设空间中经验误差最小的学习器，$f^* \in H_\kappa$ 为假设空间中泛化误差最小的学习器。

证明. 该定理的证明结合使用了 Rudi 等工作[78] 中的定理 3 及定理 8，将 Nyström–LapRLS 的泛化误差界 (定理 3.20) 与 PCG 方法的泛化分析相集合，从而得到最终结果。详细证明过程在已完成工作[135] 的定理 1 的证明中给出。 □

在不使用容量假设 (假设 3.14，$\gamma = 1$)、正则化假设 (假设 3.15，$r = 1/2$) 的情况下，定理 3.21 中给出了与标准核岭回归方法相同的泛化误差界[70]。定理 3.21 说明 $s = \mathcal{O}(\sqrt{n})$ 个 Nyström 采样点，$\mathcal{O}(\log n)$ 次迭代可以达到 $\mathcal{O}(1/\sqrt{n})$ 的泛化误差收敛率，该泛化误差收敛率与传统 LapRLS 的泛化误差收敛率相同[228]。

3.5 本章小结

如图 3.4 所示，本文针对大规模半监督的核方法模型选择泛化理论研究已取得突破性进展，已完成工作从如下三个方面递进展开进行:

1. 通过核矩阵谱分析，建立基于的谱度量的二分类核方法泛化误差理论。已完成工作中给出了 LSSVM 泛化误差界、SVM 泛化误差界[206]。

2. 通过核函数谱分析，建立基于局部 Rademacher 复杂度的多输出泛化误差理论。该部分工作主要由三部分组成：首次将局部 Rademacher 复杂度引入到核多分类中，获得收敛率为 $\mathcal{O}\left(\frac{\log K}{n}\right)$ 的多分类泛化误差界[205]；建立假设空间上局部 Rademacher 复杂度、假设空间上局部 Rademacher 复杂度的关联，使用无标签数据提升学习模型数据依赖的泛化误差界，获得收敛率为 $\mathcal{O}\left(\frac{K}{\sqrt{n+u}}+\frac{1}{n}\right)$ 的线性多分类泛化误差界[207]；整合之前工作，将局部 Rademacher 复杂度泛化误差界扩展到半监督多输出问题 (vector-valued learning)，并为核学习器、线性学习器建立收敛率为 $\mathcal{O}\left(\frac{1}{\sqrt{n+u}}+\frac{1}{n}\right)$ 的多输出泛化误差界[107,108,208]。

3. 基于积分算子理论，使用容量假设、正则化假设，为分布式、随机特征相结合的大规模半监督核岭回归 (KRR) 提供最优泛化误差理论保证[188]；基于积分算子理论，为 Nyström 采样、PCG 相结合的大规模半监督核岭回归 (LapRLS) 提供泛化理论保证[135]。

第4章　大规模半监督的核方法模型选择准则研究

基于第 3 章中建立的半监督核方法的泛化误差界，通过最小化泛化误差界，来指导核方法模型选择准则的制定。本章提出三种核方法核方法模型选择准则：最大化谱度量[206]、最小化核矩阵尾部特征值[205]、反向传播更新核超参数[107]。

4.1　最大化谱度量

将 LSSVM 泛化误差界（定理 3.3）、SVM 泛化误差界（定理 3.5）统一写为

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \leq 1 - c_6 \cdot \mathrm{SM}(\kappa, \varphi) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \tag{4.1}$$

其中，$c_6 \in \mathbb{R}_+$ 为常数。为获得更好的泛化性能，最小化 (4.1) 中的泛化误差界 $\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*)$，可得使用最大化谱度量作为二分类核方法模型选择准则

$$\arg\max_{\kappa \in \mathcal{K}} \ \mathrm{SM}(\kappa, \varphi).$$

其中，$\mathcal{K}$ 为候选的核函数结合。同时，为方便谱度量的计算，使用如下的高阶形式谱度量。

定义 4.1 (高阶形式谱度量)**.** 对核矩阵特征值使用高阶形式 $\varphi(\lambda_i) = \lambda_i^r, r \geq 1$。高阶形式谱度量具体定义为

$$\mathrm{SM}(\kappa, \varphi) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^r \langle \boldsymbol{y}, \mathbf{v}_i \rangle^2 = \frac{1}{n}\boldsymbol{y}^{\mathrm{T}}\left(\sum_{i=1}^{n} \lambda_i^r \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}\right)\boldsymbol{y} = \frac{1}{n}\boldsymbol{y}^{\mathrm{T}}\mathbf{N}^r\boldsymbol{y}.$$

定义 4.1 中的谱度量的时间复杂度为 $\mathcal{O}(n^2)$，相对于其他核方法模型选择方法计算效率较高。因此，使用如下准则选取核函数

$$\arg\max_{\kappa \in \mathcal{K}} \mathrm{SM}(\kappa, t^r) = \frac{1}{n}\boldsymbol{y}^{\mathrm{T}}\mathbf{N}^r\boldsymbol{y}$$

其中，$\mathcal{K}$ 为候选的核函数结合。为避免正负样本不均衡问题，使用加权后的最大化谱度量作为核函数选择准则：

$$\arg\max_{\kappa \in \mathcal{K}} \overline{\mathrm{SM}}(\kappa, t^r) = \frac{1}{n}\bar{\boldsymbol{y}}^{\mathrm{T}}\mathbf{N}^r\bar{\boldsymbol{y}}, \tag{4.2}$$

其中 $\bar{y}_+ = \frac{n}{n_+}$、$\bar{y}_- = -\frac{n}{n_-}$，$n_+$、$n_-$ 分别为正负样本个数。

表 4.1 最大化谱度量与其他核函数选择方法对比

模型选择准则	时间复杂度	理论保证
近似交叉验证 (CV)[16]	$\mathcal{O}(n^3)$	有
核对齐 (KTA)[229]	$\mathcal{O}(n^2)$	无
中心化核对齐 (CKTA)[230]	$\mathcal{O}(n^2)$	无
FSM[23]	$\mathcal{O}(n^2)$	无
ER[231]	$\mathcal{O}(n^3)$	有
谱度量 (SM)	$\mathcal{O}(n^2)$	有

如表 4.1 中所示，近似交叉验证方法、ER 方法虽然有理论保证，但计算复杂度很高 $\mathcal{O}(n^3)$，不适用于大规模数据上的核函数选择；而 KTA、CKTA、FSM 等方法，虽然计算效率较高，但无法从理论上保障算法的泛化性能；仅有最大化谱度量 (SM) 兼顾了计算效率、理论保证。

4.1.1 最大化谱度量的特例

本节介绍两个常用的方法：最小化图割 (minimum graph cut)、最大化均值差异 (maximum mean discrepancy, MMD)，同时证明它们是最大化谱度量的特例。

4.1.1.1 与最小化图割的关联

在图论中，图割用于衡量图中不同分块的相似性[232]。而核函数 $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 天然地反映了输入样本 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$的相似性，因此使用核矩阵 $\mathbf{K}$ 作为相似度矩阵构造图。此时，正则化图割 (normalized graph cut, Ncut)[232] 可以写为

$$\mathrm{Ncut}(\kappa) = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{L}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\mathbf{D}\boldsymbol{y}},$$

其中Laplacian 矩阵为 $\mathbf{L} = \mathbf{D} - \mathbf{K}$。 $\mathbf{D}$ 为对角矩阵，其对角元素定义为 $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{K}_{ij}$。如果正负样本个数是均衡的，即 $n_+ = n_-$，可得

$$\boldsymbol{y}^{\mathrm{T}}\mathbf{D}\boldsymbol{y} = \sum_{i=1}^{n} \mathbf{D}_{ii} = \sum_{i,j=1}^{n} \mathbf{K}_{ij} = |\mathbf{K}|_1.$$

令 $\varphi(t) = t$，将 Ncut 与谱度量关联起来

$$\mathrm{Ncut}(\kappa) = 1 - \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{K}\boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\mathbf{D}\boldsymbol{y}} = 1 - \frac{\bar{\boldsymbol{y}}^{\mathrm{T}}\mathbf{K}\bar{\boldsymbol{y}}}{2|\mathbf{K}|_1} = 1 - \frac{n}{2} \cdot \overline{\mathrm{SM}}(\kappa, t),$$

从上面的等式中可得，最小化图割等价于最大谱度量。

4.1.2　与最大均值差异的关联

均值差异 (mean discrepancy, MD) 用于衡量两个分布之间的差异[233]。再生核希尔伯特空间 $\mathcal{H}$ 上的均值差异可以写为

$$\mathrm{MD}(\mathcal{H}) = \left\| \frac{1}{n_+} \sum_{i}^{n_+} \phi(\boldsymbol{x}_i) - \frac{1}{n_-} \sum_{j}^{n_-} \phi(\boldsymbol{x}_j) \right\|_{\mathcal{H}},$$

其中 $\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$。

如果核矩阵 $\mathbf{K}$ 为正则化矩阵，即 $\mathbf{K} = \mathbf{N}$。令 $\varphi(t) = t$，易得

$$\begin{aligned}
&\overline{\mathrm{SM}}(\kappa, t) = \frac{1}{n} \bar{\mathbf{y}}^{\mathrm{T}} \mathbf{N} \bar{\mathbf{y}} = \frac{1}{n} \bar{\mathbf{y}}^{\mathrm{T}} \mathbf{K} \bar{\mathbf{y}} \\
&= \frac{1}{n_+^2} \sum_{i,j}^{n_+} \mathbf{K}_{ij} + \frac{1}{n_-^2} \sum_{i,j}^{n_-} \mathbf{K}_{ij} - \frac{2}{n_- n_+} \sum_{i}^{n_+} \sum_{j}^{n_-} \mathbf{K}_{ij} \\
&= \left\| \frac{1}{n_+} \sum_{i}^{n_+} \phi(\boldsymbol{x}_i) - \frac{1}{n_-} \sum_{j}^{n_-} \phi(\boldsymbol{x}_j) \right\|_{\mathcal{H}}.
\end{aligned}$$

上面的等式说明均值差异 $\mathrm{MD}(\kappa)$ 是谱度量 $\overline{\mathrm{SM}}(K, \varphi)$ 的特例。

注. 当使用 $\varphi(t) = t$ 时候，图割、均值差异是谱度量的特例。但在核方法模型选择准则推导、实际使用中，使用 $\varphi(t) = t^r, r \geq 1$，因为该形式计算效率较高，同时能够达到去噪的目的。

注. 不同于传统核方法模型选择在候选核函数中选取一个核函数，多核学习 (multiple kernel learning, MKL) 将候选核函数组合起来，而核模型选择过程变为学习核函数组合参数[91,234–236]。最大化谱度量可以应用到多核学习中：

$$\arg\min_{\boldsymbol{\mu}} \ \overline{\mathrm{MS}}(\kappa_{\boldsymbol{\mu}}, t^r) \quad \text{s.t.} \quad \|\boldsymbol{\mu}\|_p = 1, \ \boldsymbol{\mu} \geq 0,$$

其中 $\kappa_{\boldsymbol{\mu}} = \sum_{i=1}^{k} \mu_i \kappa_i$。利用梯度算法可以有效地解决上述优化问题。

4.2　最小化核矩阵尾部特征值

由泛化理论分析中定理 3.10、推论 3.11可知，如果核函数对应假设空间的局部 Rademacher 复杂度更小，则该核学习器能够获得更紧的泛化误差界。对局部 Rademacher 复杂度的估计 (3.27) 可以缩写为

$$\mathcal{R}(H_r) \leq 2 \sqrt{\frac{1}{n+u} \min_{\theta \geq 0} \left(c_7 \theta + \sum_{j > \theta} \lambda_j \right)}.$$

其中，c_7 为常数。在给定截断点 θ 的情况下，局部 Rademacher 复杂度主要依赖于积分算子的尾部特征值之和 $\sum_{j>\theta}\lambda_j$。而积分算子的经验形式即为核矩阵，所以可以使用核矩阵的尾部特征值之和界定经验形式的局部 Rademacher 复杂度。

对于在候选核函数中选取一个核函数的情况，可以将最小化核矩阵尾部特征值作为核方法模型选择准则

$$\arg\min_{\kappa\in\mathcal{K}} \sum_{j>\theta}\lambda_j(\mathbf{K}), \tag{4.3}$$

其中，$\mathbf{K}$ 为核函数对应的核矩阵，$\lambda_j(\mathbf{K})$ 为核函数的第 j 个降序特征值。

上述单核选择准则 (4.3) 存在两个不足：（1）最小化核矩阵的尾部特征值之和只与样本输入 $\{\boldsymbol{x}_i\}_{i=1}^{n+u}$ 相关，但忽略了样本输出 $\{\boldsymbol{y}_i\}_{i=1}^{n}$ 中的信息，因此选择出来的核函数存在一定局限性；（2）该模型选择准则割裂了核函数选择、模型训练的过程，而端对端形式同时学习核函数、模型参数往往性能更优。

基于以上不足，考虑将最小化尾部特征值之和作为学习目标函数的一部分。但对于单核方法来说，其尾部特征值是固定的，对指导核函数选择没有意义，无法使用该端对端形式进行学习。然而，多核学习 (MKL) 对应的尾部特征值之和是不固定的，与多核组合系数相关。因此，下面介绍出端对端的核方法模型选择方法：将最小化尾部特征值之和，同时学习多核组合系数、核模型参数。并可以使用有标签数据、无标签数据上定义的核矩阵 $\mathbf{K}=\{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)\}_{i,j=1}^{n+u}$，从而使用无标签数据减小 Rademacher 复杂度。

多核学习使用核函数为

$$\kappa_{\boldsymbol{\mu}}(\boldsymbol{x},\boldsymbol{x}')=\sum_{m=1}^{M}\mu_m\kappa_m(\boldsymbol{x},\boldsymbol{x}'),$$

其中，$\boldsymbol{\mu}$ 为多核组合系数，每个核函数存在 $\kappa_m(\boldsymbol{x},\boldsymbol{x})=\langle\phi_m(\boldsymbol{x}),\phi_m(\boldsymbol{x}')\rangle$。多核组合的核函数为 $\kappa_{\boldsymbol{\mu}}(\boldsymbol{x},\boldsymbol{x}')=\langle\phi_{\boldsymbol{\mu}}(\boldsymbol{x}),\phi_{\boldsymbol{\mu}}(\boldsymbol{x}')\rangle$，其诱导的隐式特征映射为 $\phi_{\boldsymbol{\mu}}(\boldsymbol{x})=\left[\sqrt{\mu_1}\phi_1(\boldsymbol{x}),\cdots,\sqrt{\mu_M}\phi_M(\boldsymbol{x})\right]^{\top}$。将局部假设空间 (3.18) 重写为多核形式

$$H_{\boldsymbol{\mu}}=\{f\,\big|\,f(\boldsymbol{x})=\langle\boldsymbol{W},\phi_{\boldsymbol{\mu}}(\boldsymbol{x})\rangle,\ \|\boldsymbol{W}\|_{2,p}\le 1,\ \mathbb{E}\,(\ell_f-\ell_{f^*})^2\le r\}.$$

4.2.1 多核凸组合方法 (Conv-MKL)

多核方法对应假设空间 $H_{\boldsymbol{\mu}}$ 的全局 Rademacher 复杂度可以由核矩阵 $\mathbf{K}_{\boldsymbol{\mu}}=\sum_{m=1}^{M}\mathbf{K}_m$ 的特征值之和（核矩阵的秩）界定。为最小化 $H_{\boldsymbol{\mu}}$ 全局 Rademacher 复杂度，常直接对核矩阵的迹进行限定[95,237]，如 $\mathrm{Tr}(\mathbf{K}_{\boldsymbol{\mu}})\le 1$。

根据泛化理论分析结果（定理 3.10、推论 3.11）可知，局部 Rademacher 复杂度由核矩阵的尾部特征值之和界定，得到的泛化误差收敛率比使用全部特征值之和（核矩阵的迹）的泛化误差收敛率快很多。类似于对核矩阵的迹进行限定，使用如下形式对尾部特征值之和进行限定：

$$H_1 = \left\{ f \in H_{\boldsymbol{\mu}} : \sum_{j>\theta} \lambda_j(\mathbf{K}_{\boldsymbol{\mu}}) \le 1 \right\},$$

其中$\lambda_j(\mathbf{K}_{\boldsymbol{\mu}})$是组合后核矩阵 $\mathbf{K}_{\boldsymbol{\mu}}$ 的第 j 大的特征值，θ 为特征值截断点（截断点之后为尾部特征值）。值得注意的是，尾部特征值之和是核矩阵迹与前 θ 个大特征值之和的差:

$$\sum_{j>\theta} \lambda_j(\mathbf{K}_{\boldsymbol{\mu}}) = \mathrm{Tr}(\mathbf{K}_{\boldsymbol{\mu}}) - \sum_{j=1}^{\theta} \lambda_j(\mathbf{K}_{\boldsymbol{\mu}}),$$

因此，基核的核矩阵尾部特征值之和可以在 $O((n+u)^2\theta)$ 的时间内计算得出。容易看出假设空间 H_1 是非凸的，同时又由:

$$\sum_{m=1}^{M} \mu_m \sum_{j>\theta} \lambda_j(\mathbf{K}_m) = \sum_{m=1}^{M} \mu_m / \|\boldsymbol{\mu}\|_1 \sum_{j>\theta} \lambda_j(\|\boldsymbol{\mu}\|_1 \mathbf{K}_m) \le \sum_{j>\theta} \lambda_j\left(\mathbf{K}_{\boldsymbol{\mu}}\right).$$

构造出凸的假设空间 H_2:

$$H_2 = \left\{ f \in H_{\boldsymbol{\mu}} : \sum_{m=1}^{M} \mu_m \sum_{j>\theta} \lambda_j(\mathbf{K}_m) \le 1 \right\}.$$

对所有基核的核矩阵进行规范化

$$\widetilde{\mathbf{K}}_m = \left(\sum_{j>\theta} \lambda_j(\mathbf{K}_m) \right)^{-1} \mathbf{K}_m$$

因此，规范化基核核函数、隐式特征映射为

$$\widetilde{\kappa}_{\boldsymbol{\mu}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{m=1}^{M} \frac{\mu_m \kappa_m(\boldsymbol{x}, \boldsymbol{x}')}{\sum_{j>\theta} \lambda_j(\mathbf{K}_m)},$$

$$\widetilde{\phi}_{\boldsymbol{\mu}}(\boldsymbol{x}) = \left[\sqrt{\frac{\mu_1}{\sum_{j>\theta} \lambda_j(\mathbf{K}_1)}} \phi_1(\boldsymbol{x}), \cdots, \sqrt{\frac{\mu_M}{\sum_{j>\theta} \lambda_j(\mathbf{K}_M)}} \phi_M(\boldsymbol{x}) \right].$$

可以将假设空间 H_2 写作

$$H_2 = \left\{ f(\boldsymbol{x}) = \langle \boldsymbol{W}, \widetilde{\phi}_{\boldsymbol{\mu}}(\boldsymbol{x}) \rangle, \|\boldsymbol{W}\|_{2,p} \le 1, \boldsymbol{\mu} \ge 0, \|\boldsymbol{\mu}\|_1 \le 1 \right\},$$

基于正则化核矩阵的假设空间 H_2，使用满足 $\boldsymbol{\mu} \ge 0, \|\boldsymbol{\mu}\|_1 \le 1$ 的多核组合系数，此时的多核学习器如算法 1 所示。

算法 1 多核凸组合方法 (Conv-MKL)

输入: M 个基核矩阵 $\mathbf{K}_1, \ldots, \mathbf{K}_M$，尾部特征值截断点 θ。

for $m = 1$ **to** M **do**

计算基核矩阵的尾部特征值: $r_m = \sum_{j>\theta} \lambda_j(\mathbf{K}_m)$。

正则化基核矩阵: $\widetilde{\mathbf{K}}_m = \mathbf{K}_m / r_m$。

end for

将正则化后的基核矩阵 $\widetilde{\mathbf{K}}_m, m = 1, \ldots, M$ 应用到任意 ℓ_p-范数多核学习器中。

4.2.2 多核学习方法 (SMSD-MKL)

多核凸组合方法中，令尾部特征值和满足 $\sum_{m=1}^{M} \mu_m \sum_{j>\theta} \lambda_j(\mathbf{K}_m) \leq 1$，而不是最小化尾部特征值之和，同时多核凸组合方法中没有对组合系数进行学习。

本节将最小化尾部特征值之和与经验损失最小化 (ERM) 一起作为学习目标，以端到端的形式同时学习多核组合系数 $\boldsymbol{\mu}$、核学习器。在假设空间 H_1 使用经验损失最小，并最小化尾部特征值之和获得更小的局部 Rademacher 复杂度:

$$\min_{\boldsymbol{W},\boldsymbol{\mu}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i)}_{C(\boldsymbol{W})} + \underbrace{\frac{\alpha}{2}\|\boldsymbol{W}\|_{2,p}^2 + \beta \sum_{m=1}^{M} \mu_m r_m}_{\Omega(\boldsymbol{W})}, \tag{4.4}$$

其中，损失函数为 $\ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i) = \left| 1 - \left(\boldsymbol{y}_i^\top f(\boldsymbol{x}_i) - \max_{y \neq \boldsymbol{y}_i} \boldsymbol{y}^\top f(\boldsymbol{x}_i) \right) \right|_+$、第 m 个基核矩阵的尾部特征值之和为 $r_m = \sum_{j>\theta} \lambda_j(\mathbf{K}_m), \forall\, m = 1, \ldots, M$。

梯度下降算法求解优化目标，需要对目标函数中各项求解一阶梯度，但由于最小化核矩阵的尾部特征值之和在很多情况下不可微，所以无法直接使用随机梯度下降算法进行求解。而合页损失只有少数情况是不可微的，所以使用次微分求解合页损失的梯度。

基于对偶梯度下降框架[238,239]，使用次微分对偶梯度下降算法 (stochastic mirror and sub-gradient descent, SMSD-MKL) 求解多核学习问题 (4.4)，从而求得多核组合系数 $\boldsymbol{\mu}$ 及核学习器 $f(\boldsymbol{x})$。使用对偶梯度下降算法求解分为两个步骤

- 使用 $C(\boldsymbol{W})$ 更新对偶权重 $\boldsymbol{\theta}$.
- 根据 $\Omega(\boldsymbol{W})$ 的费舍尔对偶(Fenchel dual)更新原始权重 $\boldsymbol{W}$、多核组合系数 $\boldsymbol{\mu}$.

详细的求解过程在第 5 章中的算法 2 进行了介绍。

4.3　反向传播更新核超参数

使用最大化谱度量作为核方法模型选择包括两个步骤：首先从核函数候选集中选取使得谱度量最大的核函数，再基于选择出来的核函数学习核学习器。该方法割裂了核函数选择、训练核学习器两个过程。

而使用最小化核矩阵尾部特征值之和作为核方法模型选择准则，虽然能够以端对端地形式同时学习核函数、核学习器，但其实际上学到核函数为多核组合的形式，计算效率较低。

借鉴神经网络中大获成功的反向传播方法，本节使用反向传播更新核超参数，该方法能够根据优化目标同时学习核超参数、模型参数，同时降低了推论 3.11 中的核方法泛化误差界。

4.3.1　自动谱核学习

4.3.1.1　平稳谱核

首先介绍常见的平稳谱核 (stationary spectral kernel)，又称为平移不变核 (shift-invariant kernel)，包括高斯核 $\kappa(\boldsymbol{x},\boldsymbol{x}') = \exp(-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2/2\sigma^2)$、拉普拉斯核 $\kappa(\boldsymbol{x},\boldsymbol{x}') = \exp(-\|\boldsymbol{x}-\boldsymbol{x}'\|_1/2\sigma^2)$ 等常用核函数。平稳谱核仅依赖于样本之间的距离（相似度）$\tau = \boldsymbol{x}-\boldsymbol{x}'$，平稳谱核的核函数定义为 $\kappa(\boldsymbol{x},\boldsymbol{x}') = \kappa(\boldsymbol{x}-\boldsymbol{x}') = \kappa(\tau)$。

通过傅里叶逆变换 (inverse Fourier transform)， Bochner 定理说明平稳谱核核函数 $\kappa(\tau)$ 由它对应的谱密度 $s(\boldsymbol{\omega})$ 所决定[240]。

引理 4.1 (Bochner 定理[31])**.** *当且仅当平稳谱核 $\kappa(\boldsymbol{x},\boldsymbol{x}') = \kappa(\boldsymbol{x}-\boldsymbol{x}')$ 能够表示为如下形式时*

$$\kappa(\boldsymbol{x},\boldsymbol{x}') = \int_{\mathcal{X}} e^{i\boldsymbol{\omega}^\top(\boldsymbol{x}-\boldsymbol{x}')} s(\boldsymbol{\omega}) d\boldsymbol{\omega}, \tag{4.5}$$

定义在输入空间 $\mathcal{X}$ 上平稳谱核 $\kappa(\boldsymbol{x},\boldsymbol{x}')$ 的是正定的 (positive definite)。其中 $s(\boldsymbol{\omega})$ 是 $\kappa(\boldsymbol{x},\boldsymbol{x}')$ 对应的非负谱密度。

从 Bochner 定理中可以看出，平稳谱核 $\kappa(\boldsymbol{x},\boldsymbol{x}')$ 是由谱密度 $s(\boldsymbol{\omega})$ 唯一确定的，因此可以通过更新谱密度 $s(\boldsymbol{\omega})$ 来学习平稳谱核。表 4.2 中列举了常见的平稳谱核 $\kappa(\tau)$ 以及其对应的谱密度 $s(\boldsymbol{\omega})$。比如，超参数为 σ 的高斯核由概率密度函数为 $\mathcal{N}(0, 1/\sigma^2)$ 的高斯分布决定。

表 4.2 平稳谱核及其对应谱密度

平稳谱核	$\kappa(\tau)$	$s(\omega)$	对应分布
高斯核	$\exp\left(-\frac{\|\tau\|_2^2}{2\sigma^2}\right)$	$\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{\sigma^2\|\omega\|_2^2}{2}\right)$	正太分布
拉普拉斯核	$\exp\left(-\frac{\|\tau\|_1}{\sigma}\right)$	$\frac{\sigma}{\pi(1+\sigma^2\omega^2)}$	柯西分布
柯西核	$\Pi_d \frac{2}{1+\tau_d^2}$	$\exp(-\|\tau\|_1)$	指数分布

基于 Bochner 定理 (4.5)，Rahimi 和 Recht[31] 提出使用蒙特卡洛采样 (Monte Carlo sampling) 生成随机傅里叶特征近似平稳谱核 $\kappa(\boldsymbol{x},\boldsymbol{x}') \approx \langle \phi_M(\boldsymbol{x}), \phi_M(\boldsymbol{x}')\rangle$。其中，随机傅里叶特征可以定义为

$$\phi_M(\boldsymbol{x}) = \sqrt{\frac{2}{M}}\cos(\boldsymbol{\Omega}^\top \boldsymbol{x} + \boldsymbol{b}), \tag{4.6}$$

其中，频率矩阵 $\boldsymbol{\Omega} = \{\omega_1, \omega_2, \cdots, \omega_M\}$ 是基于谱密度 $s(\omega)$ 采样得来。相位向量 $\boldsymbol{b} = \{b_1, b_2, \cdots, b_M\}$ 从均匀分布 $U(0, 2\pi)$ 重复采样 M 次得到。

4.3.1.2 非平稳谱核

虽然平稳谱核得到了广泛应用，但平稳谱核仅依赖于样本间距离 $\tau = \boldsymbol{x} - \boldsymbol{x}'$ 而忽略了输入样本 $\boldsymbol{x}$ 本身所携带的信息 (input-independent)。同时，很多核函数不属于平稳谱核，比如线性核 $\kappa(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}^\top\boldsymbol{x}'$、多项式核 $\kappa(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}^\top\boldsymbol{x}'+1)^r$ 等。而非平稳谱核 (non-stationary spectral kernels) 依赖于输入样本，并通过 Yaglom 定理给出了非平稳谱核与谱密度之间的关联[101]。

定理 4.2 (Yaglom 定理[101])**.** *当且仅当一般核函数 $\kappa(\boldsymbol{x},\boldsymbol{x}')$ 满足以下形式*

$$\kappa(\boldsymbol{x},\boldsymbol{x}') = \int_{\mathbb{R}^d\times\mathbb{R}^d} e^{i(\omega^\top\boldsymbol{x} - \omega'^\top\boldsymbol{x}')}\mu(d\omega, d\omega'), \tag{4.7}$$

该核函数在输入空间 $\mathcal{X}$ 是正定的。其中 $\mu(d\omega, d\omega')$ 为与半正定 (positive semi definite, PSD) 谱度量 $s(\omega,\omega')$ 相关的 Lebesgue-Stieltjes 度量。

Yaglom 定理将一般核函数 $\kappa(\boldsymbol{x},\boldsymbol{x}')$ 通过频率 ω,ω' 与半正定的谱密度 $s(\omega,\omega')$ 关联起来。同时，平稳谱核 (Bochner 定理) 是非平稳谱核 (Yaglom 定理) 的特例：当谱度量 $s(\omega,\omega')$ 集中在对角线上 $\omega = \omega'$，非平稳谱核退化为平稳谱核。

4.3.1.3　构造谱核学习框架

为使得 (4.7) 中的谱密度是半正定的，构造谱核使得其满足 $s(\boldsymbol{\omega},\boldsymbol{\omega}')=s(\boldsymbol{\omega}',\boldsymbol{\omega})$，同时引入对角线谱度量 $s(\boldsymbol{\omega},\boldsymbol{\omega}),s(\boldsymbol{\omega}',\boldsymbol{\omega}')$，最终将谱核构造为[101,102]

$$\begin{aligned}\kappa(\boldsymbol{x},\boldsymbol{x}') &= \int_{\mathcal{X}\times\mathcal{X}} \mathcal{E}_{\omega,\omega'}(\boldsymbol{x},\boldsymbol{x}')\mu(d\boldsymbol{\omega},d\boldsymbol{\omega}'),\\ \mathcal{E}_{\omega,\omega'}(\boldsymbol{x},\boldsymbol{x}') &= \frac{1}{4}\left[e^{i(\boldsymbol{\omega}^\top\boldsymbol{x}-\boldsymbol{\omega}'^\top\boldsymbol{x}')} + e^{i(\boldsymbol{\omega}'^\top\boldsymbol{x}-\boldsymbol{\omega}^\top\boldsymbol{x}')} + e^{i\boldsymbol{\omega}^\top(\boldsymbol{x}-\boldsymbol{x}')} + e^{i\boldsymbol{\omega}'^\top(\boldsymbol{x}-\boldsymbol{x}')}\right].\end{aligned} \tag{4.8}$$

与构造平稳谱核的随机特征 (4.6) 相似，使用蒙特卡洛采样构造随机特征来近似 (4.8) 中的非平稳谱核：

$$\begin{aligned}k(\boldsymbol{x},\boldsymbol{x}') &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathcal{E}_{\omega,\omega'}(\boldsymbol{x},\boldsymbol{x}')\mu(d\boldsymbol{\omega},d\boldsymbol{\omega}')\\ &= \mathbb{E}_{\boldsymbol{\omega},\boldsymbol{\omega}'\sim s(\boldsymbol{\omega},\boldsymbol{\omega}')}\left[\mathcal{E}_{\omega,\omega'}(\boldsymbol{x},\boldsymbol{x}')\right]\\ &= \mathbb{E}_{\boldsymbol{\omega},\boldsymbol{\omega}'\sim s(\boldsymbol{\omega},\boldsymbol{\omega}')}\frac{1}{4}\big[\cos(\boldsymbol{\omega}^\top\boldsymbol{x}-\boldsymbol{\omega}'^\top\boldsymbol{x}')+\cos(\boldsymbol{\omega}'^\top\boldsymbol{x}-\boldsymbol{\omega}^\top\boldsymbol{x}')\\ &\quad+\cos(\boldsymbol{\omega}^\top\boldsymbol{x}-\boldsymbol{\omega}^\top\boldsymbol{x}')+\cos(\boldsymbol{\omega}'^\top\boldsymbol{x}-\boldsymbol{\omega}'^\top\boldsymbol{x}')\big]\\ &\approx \frac{1}{4M}\sum_{i=1}^{M}\Big[\cos(\boldsymbol{\omega}_i^\top\boldsymbol{x}-\boldsymbol{\omega}_i'^\top\boldsymbol{x}')+\cos(\boldsymbol{\omega}_i'^\top\boldsymbol{x}-\boldsymbol{\omega}_i^\top\boldsymbol{x}')\\ &\quad+\cos(\boldsymbol{\omega}_i^\top\boldsymbol{x}-\boldsymbol{\omega}_i^\top\boldsymbol{x}')+\cos(\boldsymbol{\omega}_i'^\top\boldsymbol{x}-\boldsymbol{\omega}_i'^\top\boldsymbol{x}')\Big]\\ &= \langle\phi_M(\boldsymbol{x}),\phi_M(\boldsymbol{x}')\rangle\end{aligned}$$

其中 $(\boldsymbol{\omega}_i,\boldsymbol{\omega}_i')_{i=1}^M \overset{\text{i.i.d.}}{\sim} s(\boldsymbol{\omega},\boldsymbol{\omega}')$，$M$ 为蒙特卡洛采样次数。使用蒙特卡洛采样构造的非平稳谱核对应随机傅里叶特征为

$$\phi_M(\boldsymbol{x})=\frac{1}{\sqrt{4M}}\begin{bmatrix}\cos(\boldsymbol{\Omega}^\top\boldsymbol{x})+\cos(\boldsymbol{\Omega}'^\top\boldsymbol{x})\\ \sin(\boldsymbol{\Omega}^\top\boldsymbol{x})+\sin(\boldsymbol{\Omega}'^\top\boldsymbol{x})\end{bmatrix},$$

其中随机特征映射为 $\phi_M:\mathbb{R}^d\to\mathbb{R}^{2M}$，随机特征维度为 $2M$。为降低计算开销，使用与上面随机特征等价的 M 维的非平稳谱核的随机傅里叶特征为

$$\phi_M(\boldsymbol{x})=\frac{1}{\sqrt{2M}}\left[\cos(\boldsymbol{\Omega}^\top\boldsymbol{x}+\boldsymbol{b})+\cos(\boldsymbol{\Omega}'^\top\boldsymbol{x}+\boldsymbol{b}')\right], \tag{4.9}$$

其中频率矩阵 $\boldsymbol{\Omega},\boldsymbol{\Omega}'\in\mathbb{R}^{d\times M},\boldsymbol{\Omega}=\{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_M\},\boldsymbol{\Omega}'=\{\boldsymbol{\omega}_1',\boldsymbol{\omega}_2',\cdots,\boldsymbol{\omega}_M'\}$。成对频率 $\{(\boldsymbol{\omega}_i,\boldsymbol{\omega}_i')\}_{i=1}^M$ 通过蒙特卡洛方法采样于谱密度 $s(\boldsymbol{\omega},\boldsymbol{\omega}')$。偏置向量 $\boldsymbol{b},\boldsymbol{b}'$ 独立同分布地采样于均匀分布 $U(0,2\pi)$，并重复 M 次数。

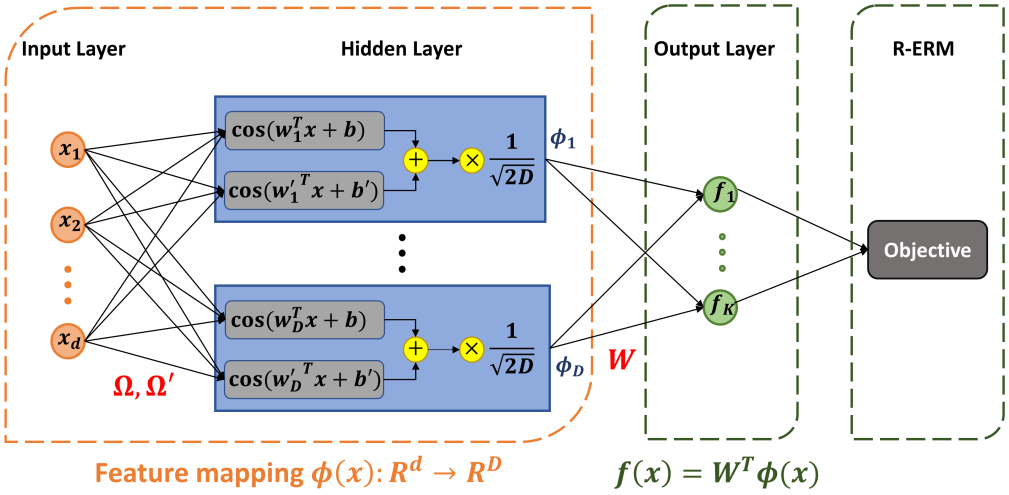

图 4.1 自动谱核学习 (ASKL) 框架

使用随机特征映射 $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$，将定义 3.1 中假设空间 3.1 重写为

$$H_\kappa = \left\{ f \mid \boldsymbol{x} \to f(\boldsymbol{x}) = \boldsymbol{W}^\top \phi_M(\boldsymbol{x}) : \|\boldsymbol{W}\|_p \le 1 \right\}.$$

其中，$\boldsymbol{W} \in \mathbb{R}^M \times \mathbb{R}^K$。特征映射 $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$ 使用 (4.9) 的形式，即

$$\phi_M(\boldsymbol{x}) = \frac{1}{\sqrt{2M}} \left[\cos(\boldsymbol{\Omega}^\top \boldsymbol{x} + \boldsymbol{b}) + \cos(\boldsymbol{\Omega}'^\top \boldsymbol{x} + \boldsymbol{b}') \right],$$

基于下一节中的泛化理论分析，使用权重矩阵的迹范数 $\|\boldsymbol{W}\|_*$、全部数据特征映射的 Frobenius 范数 $\|\phi_M(\boldsymbol{X})\|_F^2$ 能够获得更快的泛化误差收敛率。因此将 $\|\boldsymbol{W}\|_*$、$\|\phi_M(\boldsymbol{X})\|_F^2$ 作为正则化项，与经验风险最小化 (ERM) 共同作为优化目标

$$\arg\min_{\boldsymbol{W}, \boldsymbol{\Omega}, \boldsymbol{\Omega}'} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i)}_{g(\boldsymbol{W})} + \lambda_A \|\boldsymbol{W}\|_* + \lambda_B \|\phi_M(\boldsymbol{X})\|_F^2, \tag{4.10}$$

其中，$\phi_M(\boldsymbol{X}) \in \mathbb{R}^{M \times (n+u)}$ 是全部数据的特征映射，对应于核矩阵迹的近似

$$\|\phi_M(\boldsymbol{X})\|_F^2 = \sum_{i=1}^{n+u} \|\phi_M(\boldsymbol{x}_i)\|_2^2 = \sum_{i=1}^{n+u} \langle \phi_M(\boldsymbol{x}_i), \phi_M(\boldsymbol{x}_i) \rangle \approx \sum_{i=1}^{n+u} \kappa(\boldsymbol{x}, \boldsymbol{x}').$$

迹范数 $\|\boldsymbol{W}\|_*$ 对模型权重进行了规范化，而平方 Frobenius 范数 $\|\phi_M(\boldsymbol{X})\|_F^2$ 潜在地对频率进展 $\boldsymbol{\Omega}, \boldsymbol{\Omega}'$ 进行了规范化。这两个正则化项，特征是关于特征映射的罚项 $\|\phi_M(\boldsymbol{X})\|_F^2$ 在传统学习目标中很少使用。在下一节中证明了这两个正则化项、反向传播的使用，能够导致更好的泛化性能。图 4.1 展示了自动谱核学习 (automated spectral kernel learning, ASKL) 的学习框架。

在学习过程中，仅对频率矩阵 $\boldsymbol{\Omega}, \boldsymbol{\Omega}'$、权重矩阵 $\boldsymbol{W}$ 使用反向传播进行更新。使用反向传播学习频率矩阵 $\boldsymbol{\Omega}, \boldsymbol{\Omega}'$ 潜在地改变了谱密度 $s(\boldsymbol{\omega}, \boldsymbol{\omega}')$，从而使得特征映射 ϕ_M 与样本输出相关，从而获得更强大的特征表示能力。从神经网络的角度看， ASKL 是激活函数为 cosine 的单隐藏层神经网络，能够使用反向传播同时高效地更新输入层到隐藏层、隐藏层到输出层的权重；而从核方法的角度来看，ASKL 使用有限维特征（隐藏层节点）近似平移不变核，可以应用核方法的近似泛化理论。全连接的深度神经网络可以视为对多个平移不变核堆叠的近似，在每一层都使用有限维特征（有限个隐藏层节点）近似该层对应的平移不变核，核方法的泛化理论研究可以拓展到全连接的深度神经网络上。因此，ASKL 兼具了神经网络、核方法的优点，既有坚实的泛化理论基础，又能够使用反向传播高效地学习新的核函数，为核方法、神经网络的研究建立了桥梁。

4.3.2　泛化理论保证

基于定理 3.9，并使用全局 Rademacher 复杂度 ($\theta = 0$)，容易得到如下推论。

推论 4.3 (谱核的泛化误差界)**.** 令 $B = \sup_{f \in H_\kappa} \|\boldsymbol{W}\|_* < \infty$，并假设损失函数 ℓ 满足 $\ell - 2$ 范数下的 *L-Lipschitz* 连续。则以至少 $1 - \delta$ 的概率存在如下泛化误差界

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \leq c_8 \widehat{\mathcal{R}}(H_\kappa) + \mathcal{O}\left(\frac{1}{n}\right), \tag{4.11}$$

其中 c_8 为常数， $f^* \in H_\kappa$ 为泛化误差最小模型，$\widehat{f}_n \in H_\kappa$ 为经验误差最小对应模型。同时，在使用随机傅里叶特征近似非平稳谱核 (4.9) 的情况下，全局 *Rademacher* 复杂度存在如下上界

$$\begin{aligned}
&\widehat{\mathcal{R}}(H_\kappa) \\
&\leq \frac{B}{n+u} \sqrt{K \sum_{i=1}^{n+u} \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_i) \rangle} \\
&= \mathbb{E}_{(\boldsymbol{\omega}, \boldsymbol{\omega}') \sim s(\boldsymbol{\omega}, \boldsymbol{\omega}')} \ \frac{B}{n+u} \sqrt{K \sum_{i=1}^{n+u} \frac{1}{2} \left[\cos\left((\boldsymbol{\omega} - \boldsymbol{\omega}')^\top \boldsymbol{x}_i \right) + 1 \right]} \\
&\approx \frac{B}{n+u} \sqrt{\frac{K}{M} \sum_{i=1}^{n+u} \sum_{j=1}^{M} \frac{1}{2} \left[\cos\left((\boldsymbol{\omega}_j - \boldsymbol{\omega}_j')^\top \boldsymbol{x}_i \right) + 1 \right]}.
\end{aligned} \tag{4.12}$$

讨论如下：

(1) 全局 Rademacher 复杂度主要与权重矩阵迹范数 $B = \sup_{f \in H_\kappa} \|\boldsymbol{W}\|_* < \infty$、核矩阵迹 $\sum_{i=1}^{n} \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_i) \rangle$ 两项相关，因此将 $\|\boldsymbol{W}\|_*$、$\|\phi_M(\boldsymbol{X})\|_F^2$ 作为正则化项。

(2) 对于平稳谱核，核矩阵的迹范数为定值 $\sum_{i=1}^{n}\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_i)=n$。对应于 (4.12) 中 $\boldsymbol{\omega}=\boldsymbol{\omega}'$ 的情况，存在常量 $\cos\left((\boldsymbol{\omega}-\boldsymbol{\omega}')^{\top}\boldsymbol{x}_i\right)=1$。而非平稳谱核，即 $\boldsymbol{\omega}\neq\boldsymbol{\omega}'$ 在很少情况看下能够取得 $\cos\left((\boldsymbol{\omega}-\boldsymbol{\omega}')^{\top}\boldsymbol{x}_i\right)=1$，说明了非平稳谱核在泛化性能上显著优于平稳谱核。

(3) 根据优化目标，使用反向传播更新 $(\boldsymbol{\omega},\boldsymbol{\omega}')$，能够获得更适合该学习任务的概率密度分布 $s(\boldsymbol{\omega},\boldsymbol{\omega}')$，从而获得更合适的特征映射 ϕ_M。使用反向传播改变了谱密度 $s(\boldsymbol{\omega},\boldsymbol{\omega}')$，实际上改变了对应的核函数 κ，而与任务相关 (output-dependent) 的谱核避免了较差情况，往往核矩阵的迹范数更小，泛化性能更优。

(4) 传统核方法只对权重矩阵 $\boldsymbol{W}$ 进行更新，不更新谱密度 $s(\boldsymbol{\omega},\boldsymbol{\omega}')$，因此其泛化性能极大地依赖于初始的核函数（即其对应的谱密度）。而使用自动谱核学习框架，根据目标函数使用反向传播同时更新$\boldsymbol{W}$、$s(\boldsymbol{\omega},\boldsymbol{\omega}')$，改变了初始的核函数。通过反向传播更新谱密度，学得的核函数更适合学习任务。

第5章　大规模半监督的核方法模型选择算法研究

传统核方法学习所需时间、空间复杂度均不低于$\mathcal{O}(n^2)$。常见核方法如非线性核支持向量机 (SVM) 使用序列最小化 (SMO) 算法进行求解，其空间复杂度为$\mathcal{O}(n^2)$，空间复杂度为 $\mathcal{O}(dn^2)$。而核岭回归方法 (KRR) 存在如下闭式解

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda_A \mathbf{I})^{-1}\mathbf{y}$$

其中，$\mathbf{K}$ 为核函数 κ 对应的核矩阵，$\mathbf{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]^\top$ 对应于所有样本标签组成的向量。核岭回归存储核矩阵 $\mathbf{K}$ 需要内存为 $\mathcal{O}(n^2)$；而闭式解的求解涉及到大小为 $n \times n$ 的矩阵求逆，所以核岭回归的时间复杂度为 $\mathcal{O}(n^3)$。而核方法模型选择需要训练核学习器多轮（如交叉验证），因此时间复杂度会更高，如 k-折交叉验证选择核函数的时间复杂度为 $\mathcal{O}(kn^3)$。

5.1　常用的大规模核方法加速算法

如何降低核方法模型选择内存需求、提高计算效率，使得核方法模型选择适用于大规模数据成为核方法的研究热点。目前，应用于大规模核方法的方法包括分布式、低秩近似、一阶梯度优化算法三种，具体如下：

1. 基于分布式的大规模算法

分布式方法将训练数据进行分块，并在每个分块上训练局部学习器，训练过程中经过必要的通信，最终将所有局部学习器合并为全局学习器。已完成工作[188] 使用了最简单的分布式方法：分治方法 (divide and conquer)。分治方法将训练集 D 划分为 m 份, 在每个子集 D_j 上训练产生一个局部的学习器

$$\widehat{f}_{D_j}(\boldsymbol{x}) = \sum_{\boldsymbol{x}_i \in D_j} \widehat{\alpha}_{ij}\kappa(\boldsymbol{x}_i, \boldsymbol{x}), \qquad \widehat{\alpha}_j = (\mathbf{K}_{D_j} + \lambda_A \mathbf{I})^{-1}\mathbf{y}_{D_j},$$

将局部学习器平均, 构成全局学习器 (KRR-DC)

$$\widehat{f}_D = \frac{1}{m}\sum_{j=1}^{m} \widehat{f}_{D_j}.$$

基于积分算子理论，研究分块数 m 与最优泛化误差收敛率、空间复杂度、时间复杂度的关系。当分块数满足 $m \lesssim \mathcal{O}\left(n^{\frac{2r-1}{2r+\gamma}}\right)$ 时，KRR-DC 能够达到最优

泛化误差界。同时，KRR-DC 的计算效率与分块数 m 相关，其空间复杂度为 $\mathcal{O}(n^2/m^2)$、时间复杂度为 $\mathcal{O}(n^3/m^3)$。

2. **基于核方法低秩近似的大规模算法**

非线性核方法的理论时间复杂度不会低于 $\mathcal{O}(n^2)$，常用快速近似法加速非线性核方法的求解，主要包括 Nyström 采样和随机特征两种。

(1) 使用 Nyström 方法近似核矩阵[135]

$$\mathbf{K} \approx \mathbf{K}_{ns}\mathbf{K}_{ss}^{\dagger}\mathbf{K}_{ns}^{\top},$$

其中，$\mathbf{K}_{ns}$、$\mathbf{K}_{ss}$ 为全部数据上核矩阵 $\mathbf{K}$ 中的某个分块，分别对应于全部数据与采样数据的核矩阵、采样数据上的核矩阵。将核矩阵中元素采样后重新排序，将采样的 s 个样本对应元素放在前面，可得如下的核矩阵划分

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{ss} & \mathbf{K}_{21}^{\top} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}, \qquad \mathbf{K}_{ns} = \begin{bmatrix} \mathbf{K}_{ss} \\ \mathbf{K}_{21} \end{bmatrix}.$$

使用 Nyström 方法只需要存储核矩阵 $\mathbf{K}_{ns}$，因此空间复杂度为 $\mathcal{O}(ns)$；而计算过程中可以对近似形式进一步拆分，时间复杂度为 $\mathcal{O}(ns^2)$。当 Nyström 中心点数满足 $s \gtrsim \mathcal{O}\left(n^{\frac{1}{2r+\gamma}}\right)$ 时， KRR–Nyström 能够达到最优泛化误差收敛率。KRR–Nyström 的空间复杂度为 $\mathcal{O}(ns)$、时间复杂度为 $\mathcal{O}(ns^2)$。

(2) 使用随机特征 (random features) 近似核函数[107,188]

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \phi_M(\boldsymbol{x}), \phi_M(\boldsymbol{x}') \rangle,$$

其中 ϕ_M 为核函数 κ 对应的随机特征，将核方法从隐式的特征空间 $\mathcal{H}$ 通过蒙特卡洛方法近似为有限维的 $\mathbb{R}^M$，在该特征空间下使用随机特征近似的核方法的模型训练实际上是线性方法。

当随机特征数满足 $M \geq \mathcal{O}\left(n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\right)$ 时, KRR–RF 能够达到最优泛化误差收敛率 $\mathcal{O}\left(n^{\frac{-2r}{2r+\gamma}}\right)$。但在实际计算中，随机特征与输入空间为 $\mathbb{R}^M$ 的线性方法相似，空间复杂度为 $\mathcal{O}(nM)$、时间复杂度为 $\mathcal{O}(nM^2)$。

3. **基于一阶梯度的随机优化算法**

- 已完成工作[205] 使用对偶梯度下降算法求解多核多分类问题。
- 已完成工作[207,208] 使用近端梯度下降算法求解优化目标中包含不可微项（尾部奇异值之和、迹范数）的多输出问题。
- 已完成工作[135] 使用预处理共轭梯度 (PCG) 方法加速闭式解求解。

表 5.1 近似核岭回归算法对比

方法	分块数 m	随机中心数 M	空间复杂度	时间复杂度
KRR	/	/	$\mathcal{O}\left(n^2\right)$	$\mathcal{O}\left(n^3\right)$
KRR-Nyström	/	$\mathcal{O}\left(n^{\frac{1}{2r+\gamma}}\right)$	$\mathcal{O}\left(nM\right)$	$\mathcal{O}\left(nM^2\right)$
KRR-RF	/	$\mathcal{O}\left(n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\right)$	$\mathcal{O}\left(nM\right)$	$\mathcal{O}\left(nM^2\right)$
KRR-DC	$\mathcal{O}\left(n^{\frac{2r-1}{2r+\gamma}}\right)$	/	$\mathcal{O}\left(n^2/m^2\right)$	$\mathcal{O}\left(n^3/m^3\right)$
定理 3.16	$\mathcal{O}\left(n^{\frac{2r-1}{2r+\gamma}}\right)$	$\mathcal{O}\left(n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\right)$	$\mathcal{O}\left(\frac{nM}{m}\right)$	$\mathcal{O}\left(\frac{nM^2}{m}\right)$
定理 3.17	$\mathcal{O}\left(n_* n^{\frac{-\gamma-1}{2r+\gamma}}\right)$	$\mathcal{O}\left(n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}\right)$	$\mathcal{O}\left(\frac{n_* M}{m}\right)$	$\mathcal{O}\left(\frac{n_* M^2}{m}\right)$

对于 Nyström 采样，随机中心数 M 为 Nyström 采样中心点数；
对于随机特征，随机中心点数 M 为随机特征个数。

5.2　结合分治算法、随机特征的核岭回归算法

已完成工作[188] 将分治算法、随机特征两种加速手段相结合，提升大规模核岭回归的求解效率。

5.2.1　近似核方法构造

令核函数存在随机特征映射 $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$ 满足

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \phi_M(\boldsymbol{x}), \phi_M(\boldsymbol{x}') \rangle.$$

使用分治算法，将数据划分为 m 块，并为每块数据使用随机特征求解核岭回归。第 j 个数据分块 D_j 上，局部学习器为

$$\widehat{f}_{D_j}^M(\boldsymbol{x}) = \boldsymbol{W}_j^\top \phi_M(\boldsymbol{x}), \quad \text{with} \quad \boldsymbol{W}_j = (\widehat{S}_M^\top \widehat{S}_M + \lambda_A I)^{-1} \widehat{S}_M^\top \widehat{\mathbf{y}}_j, \tag{5.1}$$

其中$\lambda_A > 0$。 $\forall(\boldsymbol{x}, y) \in D_j$，存在 $\widehat{S}_M^\top = \frac{1}{\sqrt{n/m}}(\phi_M(\boldsymbol{x}_1), \cdots, \phi_M(\boldsymbol{x}_{n/m}))$ 以及 $\widehat{\mathbf{y}}_j = \frac{1}{\sqrt{n/m}}(y_1, \cdots, y_{n/m})$。将分块上的学习器进行平均后，得到全局学习器

$$\widehat{f}_D^M = \frac{1}{m} \sum_{j=1}^{m} \widehat{f}_{D_j}^M. \tag{5.2}$$

上式 (5.2) 中的学习器 $\widehat{f}_D^M$ 结合了分治算法、随机特征两种核方法加速方法，将这种组合方法称为 KRR-DC-RF。定理 3.16 为分治算法、随机特征的组合 KRR-DC-RF 提供了最优泛化理论保证。

表 5.2 近似核岭回归空间复杂度的对比

假设 3.14	假设 3.15	KRR	KRR–Nyström	KRR–RF	KRR–DC	KRR-DC-RF
$\gamma = 1$	$r = 0.5$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$
$\gamma = 0.5$	$r = 0.75$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.625})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.375})$
$\gamma = 0$	$r = 1$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n)$	$\mathcal{O}(n)$

表 5.3 近似核岭回归时间复杂度的对比

假设 3.14	假设 3.15	KRR	KRR–Nyström	KRR–RF	KRR–DC	KRR-DC-RF
$\gamma = 1$	$r = 0.5$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$
$\gamma = 0.5$	$r = 0.75$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{2.25})$	$\mathcal{O}(n^{2.25})$	$\mathcal{O}(n^2)$
$\gamma = 0$	$r = 1$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1.5})$	$\mathcal{O}(n^{1.5})$

5.2.2 相关工作对比

表 5.1 对相关近似核岭回归算法进行了对比，对比方法包括标准核岭回归 (KRR)[70]，使用 Nyström 采样近似核矩阵的核岭回归 (KRR-Nyström)[73]，使用随机特征近似核函数的核岭回归 (KRR-RF)[74]，使用分治算法求解的核岭回归 (KRR-DC)[75]，以及本文提出的分布式、随机特征相结合的方法 KRR-DC-RF (定理 3.16)、半监督方法 SKRR-DC-RF (定理 3.17)。表 5.1 中所有近似核岭回归算法均取得最优泛化收敛率 $\mathcal{O}\left(n^{-2r/(2r+\gamma)}\right)$，比较不同假设下对应的分块数 m、随机采样个数 M，以及学习算法对应的空间复杂度、时间复杂度。从表 5.1 中可以看出： (1) 随意采样点数（Nyström 采样点、随机特征采样点） M 越少，近似算法的内存需求越小、计算效率越高； (2) 分块数 m 越大，近似算法的内存需求越小、计算效率越高； (3) 由于 $(2r-1)\gamma \geq 0$，KRR–Nyström 的采样效率略高于 KRR–RF，也就是说达到相同的空间复杂度、时间复杂度、泛化误差收敛率，KRR–Nyström 需要的采样点更少。

在表 5.2、表 5.3 中，将 γ、r 量化，以 $(\gamma, r) = (1, 0.5)$、$(0.5, 0.75)$、$(0, 1)$ 三种情况为例，对比几种近似算法的空间、时间复杂度。由于所有近似方法都达到了最优泛化误差，所以在容量假设、正则化假设不同满足情况下，各个算法泛化误差率相同。由表 5.2、表 5.3 可得，使用分布式、随机特征相结合的 KRR-DC-RF 在各情况下均取得了最小的空间复杂度、时间复杂度。相比于只使用一种加速手段的近似核岭回归，KRR-DC-RF 的计算效率有明显提升。

5.3　结合 Nyström 采样、PCG 加速的半监督核岭回归

首先依次介绍最小二乘回归 (RLS)、半监督核岭回归 (LapRLS)、使用 Nyström 加速的 LapRLS (Nyström)、使用 Nyström 加速的 LapRLS (Nyström)。再对各方法的求解所需的时间、空间复杂度进行对比。

5.3.1　近似核方法构造

5.3.1.1　最小二乘回归 (RLS)

又叫核岭回归 (KRR)，定义在有标签数据集 D^l 上，存在如下闭式解

$$\widehat{f}_n(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i^* \mathbf{K}_{nn}(\mathbf{x}_i, \mathbf{x}), \quad \boldsymbol{\alpha}^* = (\mathbf{K}_{nn} + \lambda \mathbf{I})^{-1}\mathbf{y},$$

其中 $\mathbf{K}_{nn} \in \mathbb{R}^{n\times n}$ 为定义在有标签数据上的核矩阵，$\mathbf{y}$ 为对应标签组成的向量。

5.3.1.2　半监督核岭回归 (Laplacian Regularized Least Squares, LapRLS)

使用 Laplacian 正则化，将 RLS 的闭式解推广到半监督数据上

$$\widehat{f}_n = \sum_{i=1}^{n} \widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}), \quad \widehat{\boldsymbol{\alpha}} = (\mathbf{J}\mathbf{K}_{n_*n_*} + \lambda_A \mathbf{I} + \lambda_I \mathbf{L}_{n_*n_*}\mathbf{K}_{n_*n_*})^{-1}\mathbf{y}_n.$$

为方便表示，令 $n_* = n + u$ 代表全部样本个数。其中 $\mathbf{K}_{n_*n_*} \in \mathbb{R}^{n_*\times n_*}$ 为定义在全部数据上的的核矩阵，$\mathbf{L}_{n_*n_*}$ 为定义在全部数据上的 Laplacian 矩阵。对角矩阵 $\mathbf{J} = \mathrm{diag}(1, \cdots, 1, 0, \cdots, 0)$，其中前 n 个对角线元素为 1 其他为 0，$\mathbf{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n, 0, \cdots, 0]^\top$ 前 n 个元素使用 D^l 中标签，剩余对角元素使用 0 填充。

5.3.1.3　使用 Nyström 加速的 LapRLS (LapRLS–Nyström)

使用 Nyström 采样后，从 (3.58) 推导得出 Nyström–LapRLS 的解为:

$$\widehat{f}_n = \sum_{i=1}^{s} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}), \quad \boldsymbol{\alpha} = \underbrace{(\mathbf{K}_{ns}^\top \mathbf{K}_{ns} + \lambda_A \mathbf{K}_{ss} + \lambda_I \mathbf{K}_{n_*s}^\top \mathbf{L}\mathbf{K}_{n_*s})^\dagger}_{\mathbf{H}} \underbrace{\mathbf{K}_{ns}^\top \mathbf{y}}_{\mathbf{z}},$$

其中，$\mathbf{K}_{ns}$、$\mathbf{K}_{ss}$、$\mathbf{K}_{n_*s}$ 均为 $\mathbf{K}_{n_*n_*}$ 中的某个分块。同时使用 $s \times s$ 大小的矩阵块乘法避免存储大矩阵 $\mathbf{K}_{ns}$、$\mathbf{K}_{n_*s}$。

基于泛化理论保证（定理 3.20、定理 1[73]），可令 $s = \sqrt{n_*}$，此时LapRLS–Nyström 对应的空间复杂度为 $\mathcal{O}(n_*)$、时间复杂度为 $\mathcal{O}(n_*^2)$。其中，矩阵求逆的时间复杂度 $\mathcal{O}(n_*^2)$ 过高，考虑使用迭代算法求解线性系统，比如共轭梯度下降算法 (Conjugate Gradient, CG)。迭代算法的时间复杂度依赖于迭代次数 t、每次迭代中的计算复杂度，下面展开讨论：

• 求解闭式解 $\mathbf{H}\boldsymbol{\alpha} = \mathbf{z}$ 迭代次数 t 依赖于 $\mathbf{H}$ 的条件数 (Conditon number) 所决定, 条件数越小(well-conditioned), 迭代次数越少。

$$\mathrm{cond}(\mathbf{H}) = \frac{\lambda_{\max}(\mathbf{H})}{\lambda_{\min}(\mathbf{H})}, \qquad t = \mathcal{O}\big(\mathrm{cond}(\mathbf{H})\log(1/\epsilon)\big).$$

Camoriano 等证明 $t = \sqrt{s}\log s$ 的迭代次数能保证良好的泛化性能[172]。

• 每次迭代过程中，需要计算 $\mathbf{K}_{n_* s}\boldsymbol{\alpha}$，对应时间复杂度为 $\mathcal{O}(n_* s)$。

因此，令 $s = \sqrt{n_*}$ 并忽略较小的对数项，总体时间复杂度为 $\mathcal{O}(n_* s t) = \mathcal{O}(n_*^{1.75})$。

5.3.1.4 使用 PCG 加速 LapRLS–Nyström 的闭式解求解 (Nyström-PCG)

使用预处理共轭梯度 (Preconditioned Conjugate Gradient，PCG) 算法，对上面线性系统的求解进行加速

$$\mathbf{P}^{-1}\mathbf{H}\boldsymbol{\alpha} = \mathbf{P}^{-1}\mathbf{z}.$$

其中，$\mathbf{P}$ 为预处理器。预处理器 $\mathbf{P}$ 应与 $\mathbf{H}$ 尽量相近，从而降低条件数。同时又要方便计算，因此引入如下两个预处理器

• $n \leq \sqrt{n_*}$

$$\mathbf{P} = \mathbf{K}_{n_* s}^{\top}\mathbf{K}_{n_* s} + \lambda_A \mathbf{K}_{ss} + \frac{\lambda_I n^2}{s^2}\mathbf{K}_{ss}\mathbf{L}_{ss}\mathbf{K}_{ss}.$$

• $n > \sqrt{n_*}$

$$\mathbf{P} = \frac{n}{s}\mathbf{K}_{ss}^{\top}\mathbf{K}_{ss} + \lambda_A \mathbf{K}_{ss} + \frac{\lambda_I n^2}{s^2}\mathbf{K}_{ss}\mathbf{L}_{ss}\mathbf{K}_{ss}.$$

将 LapRLS–Nyström 经验学习器 $\widehat{f_n}$ 定义为使用 PCG 方法求解，并称之为 Nyström-PCG。定理 3.21 证明了使用 PCG 求解 LapRLS–Nyström 的闭式解只需要 $s = \mathcal{O}(\sqrt{n+u})$ 个 Nyström 采样中心点、$t = \mathcal{O}(\log s)$ 次迭代就能够保证 $\mathcal{O}(1/\sqrt{n})$ 的泛化误差收敛率。此时，Nyström-PCG 方法的空间复杂度为 $\mathcal{O}(n_*)$、时间复杂度为 $\mathcal{O}(n_*^{1.5})$。

5.3.2 对比方法介绍

表 5.4 对 LapRLS 相关算法的空间复杂度、时间复杂度进行了对比。其中，LapRLS–Direct、Nyström-Direct 均直接使用矩阵运算进行求解 LapRLS、使用 Nyström 加速的 LapRLS； LapRLS–CG、Nyström-CG 均使用共轭梯度下降

表 5.4 Nyström-PCG 相关算法时间、空间复杂度对比

学习器	时间复杂度	空间复杂度
RLS-Direct	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$
LapRLS-Direct	$\mathcal{O}(n_*^3)$	$\mathcal{O}(n_*^2)$
LapRLS-CG	$\mathcal{O}(n_*^{2.5})$	$\mathcal{O}(n_*^2)$
LapRLS-PCG	$\mathcal{O}(n_*^2)$	$\mathcal{O}(n_*^2)$
Nyström-Direct	$\mathcal{O}(n_*^2)$	$\mathcal{O}(n_*)$
Nyström-CG	$\mathcal{O}(n_*^{1.75})$	$\mathcal{O}(n_*)$
Nyström-PCG	$\mathcal{O}(n_*^{1.5})$	$\mathcal{O}(n_*)$

其中，n 为有标签样本数，u 为无标签样本数，$n_* = n + u$ 代表全部样本数，$u \gg n$。

算法 (CG) 进行迭代求解 LapRLS、使用 Nyström 加速的 LapRLS； LapRLS–PCG、Nyström-PCG 均使用预处理共轭梯度下降算法 (PCG) 算法进行迭代求解 LapRLS、使用 Nyström 加速的 LapRLS。从表 5.4 可以看出：(1) CG、PCG 只用来加速线性系统求解，无法减少空间复杂度，而空间复杂度的减少来自于 Nyström 采样； (2) 矩阵运算、CG、PCG 三个算法的时间复杂度依次减少，同时 Nyström 近似很大程度上减少了时间复杂度。最终，使用 Nyström、PCG 相结合的方法求解 LapRLS 问题 (Nyström-PCG)，将空间复杂度从 $\mathcal{O}(n_*^2)$ 降低到 $\mathcal{O}(n_*)$，将时间复杂度从 $\mathcal{O}(n_*^3)$ 降低到 $\mathcal{O}(n_*^{1.5})$。

5.4　基于一阶梯度的随机优化算法

5.4.1　对偶梯度下降

针对最小化核矩阵尾部特征值之和中的多核多分类的优化目标 (4.4)，使用对偶梯度下降算法在原问题上进行求解[205]。最小化核矩阵尾部特征值之和中的多核多分类的优化目标为

$$\min_{\boldsymbol{W},\boldsymbol{\mu}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}_i), y_i)}_{C(\boldsymbol{W})} + \underbrace{\frac{\alpha}{2}\|\boldsymbol{W}\|_{2,p}^2 + \beta\sum_{m=1}^{M}\mu_m r_m}_{\Omega(\boldsymbol{W})}, \tag{5.3}$$

为方便优化目标 (4.4) 的求解，引入原始权重 $\boldsymbol{W}$ 的对偶权重 $\boldsymbol{\theta}$，使用对偶梯

算法 2 随机对偶梯度下降算法求解多核多分类优化目标 (SMSD-MKL)

输入: $\alpha, \beta, \boldsymbol{r}, T$

初始化: $\boldsymbol{W}^1 = \mathbf{0}, \boldsymbol{\theta}^1 = \mathbf{0}, \boldsymbol{\mu}^1 = \mathbf{1}, q = 2\log K$

for $t = 1$ **to** T **do**

在训练数据上随机采样 $(\boldsymbol{x}^t, \boldsymbol{y}^t)$

计算对偶权重: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \partial C(\boldsymbol{W}^t)$

计算 $\nu_m^{t+1} = \|\boldsymbol{\theta}_m^{t+1}\| - t\beta r_m,\ \forall m = 1, \ldots, M$

计算多核组合系数 $\mu_m^{t+1} = \frac{\operatorname{sgn}(\nu_m^{t+1})|\nu_m^{t+1}|^{q-1}}{\alpha\|\boldsymbol{\theta}_m^{t+1}\|\|\boldsymbol{\nu}^{t+1}|_q^{q-2}},\ \forall m = 1, \ldots, M$

end for

度下降（算法 2，SMSD–MKL）进行求解。在算法 2 的实际求解中，只更新对偶权重 $\boldsymbol{\theta}$，而 $\boldsymbol{\mu}$ 的更新是通过定理 5.1 中给出的 $\boldsymbol{W}$ 与 $\boldsymbol{\mu}$ 的关联进行更新。

将优化目标 (4.4) 划分为 $C(\boldsymbol{W})$、$\Omega(\boldsymbol{W})$ 两部分：

• 对于 $C(\boldsymbol{W})$, 通过求解 $C(\boldsymbol{W})$ 的梯度，算法 2 更新对偶权重 $\boldsymbol{\theta}$。由于 $C(\boldsymbol{W})$ 使用的合页损失是不可微的，因此使用次微分 $z^t = \partial\ell(f(\boldsymbol{x}^t), y^t)$。

• 对于 $\Omega(\boldsymbol{W})$，采用 UFO-MKL 算法类似手段[239]。基于定理 5.1，算法使用 $\boldsymbol{W} = \nabla\Omega^*(\boldsymbol{\theta})$ 更新多核组合系数 $\boldsymbol{\mu}$。

实际上，算法 2 通过更新权重范数 $\|\boldsymbol{\theta}_m^{t+1}\|$、$\nu_m^{t+1}$、$\mu_m^{t+1}$ 等实数，避免了直接更新再生核希尔伯特空间 $\mathcal{H}$ 中的原始权重 $\boldsymbol{W}^{t+1}$、对偶权重 $\boldsymbol{\theta}_m^{t+1}$。而是通过更新对偶权重的范数 $\|\boldsymbol{\theta}_m^{t+1}\|$ 进行高效计算:

$$\|\boldsymbol{\theta}_m^{t+1}\|_2^2 = \|\boldsymbol{\theta}_m^t - \boldsymbol{z}_m^t\|_2^2 = \|\boldsymbol{\theta}_m^t\|_2^2 - 2\boldsymbol{\theta}_m^t \cdot \boldsymbol{z}_m^t + \|\boldsymbol{z}_m^t\|_2^2,$$

其中，$\boldsymbol{z}^t$ 为 $C(\boldsymbol{W}^t)$ 的次微分。

定理 5.1. 令

$$\boldsymbol{\nu} = \Big[\|\boldsymbol{\theta}_1\| - \beta r_1, \ldots, \|\boldsymbol{\theta}_M\| - \beta r_M\Big],$$

则 $\nabla\Omega^*(\boldsymbol{\theta})$ 的第 m 个元素为

$$\frac{\operatorname{sgn}(\nu_m)\boldsymbol{\theta}_m}{\alpha\|\boldsymbol{\theta}_m\|}\frac{|\nu_m|^{q-1}}{\|\boldsymbol{\nu}\|_q^{q-2}},$$

其中，$\mathrm{sgn}(x)$ 定义为

$$\mathrm{sgn}(x)=\begin{cases}-1, & x<0,\\ [-1,+1], & x=0,\\ +1, & x>0.\end{cases}$$

定理 5.1 建立了多核系数 $\boldsymbol{\mu}$、对偶权重 $\boldsymbol{\theta}$ 之间的联系。在证明该定理之前，首先介绍证明过程中使用到的两个引理。

引理 5.2 (连接函数 (link function)[241])**.** 令 $p\in(1,2], q=p/(p-1)$，则范数 $\|\boldsymbol{c}\|_p$、$\|\boldsymbol{c}^*\|_q$ 互为对偶。将函数 $g:\mathcal{M}\to\mathcal{M}$ 定义为

$$c_i^*=g_i(\boldsymbol{c})=\nabla_i\left(\frac{1}{2}\|\boldsymbol{c}\|_p^2\right)=\frac{sng(c_i)|c_i|^{p-1}}{\|\boldsymbol{c}\|_p^{p-2}}, i=1,\ldots,n,$$

对偶函数 g^{-1} 定义为

$$c_i=g_i^{-1}(\boldsymbol{c}^*)=\nabla_i\left(\frac{1}{2}\|\boldsymbol{c}^*\|_q^2\right)=\frac{sng(c_i^*)|c_i^*|^{q-1}}{\|\boldsymbol{c}^*\|_q^{q-2}}, i=1,\ldots,n.$$

引理 5.3 (ℓ_1-正则化闭式解[242])**.** 对于 ℓ_1 正则化，存在如下优化目标

$$\arg\min_{\boldsymbol{W}\in\mathbf{R}}\eta_t\boldsymbol{W}+\lambda_t|\boldsymbol{W}|+\frac{\gamma_t}{2}\boldsymbol{W}^2,$$

对应的最优解 $\boldsymbol{W}^*$ 为

$$\boldsymbol{W}^*=\begin{cases}\mathbf{0} & |\eta_t|\le\lambda_t,\\ -\dfrac{1}{\gamma_t}(\eta_t-\lambda_t sgn(\eta_t)) & \text{其他}.\end{cases}$$

证明. 该最小化问题是一个无约束非光滑优化问题。最优条件说明当且仅当存在 $\xi\in\partial|\mathbf{w}^*|$ 时，$\mathbf{w}^*$ 为最优解[243]。可得

$$\eta_t+\lambda_t\xi+\gamma_t\mathbf{w}^*=0.$$

最终得到 ℓ_1-正则化的闭式解。 □

定理 *5.1* 的证明. 根据标准的 Legendre-Fenchel 对偶，可得

$$\begin{aligned}\nabla\Omega^*(\boldsymbol{\theta})&=\arg\max_{\boldsymbol{W}}\boldsymbol{W}\cdot\boldsymbol{\theta}-\Omega(\boldsymbol{W})\\ &=\arg\max_{\boldsymbol{W}}\boldsymbol{W}\cdot\boldsymbol{\theta}-\frac{\alpha}{2}\|\boldsymbol{W}\|_{2,p}^2-\beta\boldsymbol{\mu}\cdot\boldsymbol{r}.\end{aligned}\tag{5.4}$$

为求得上式的优化问题，需要将上式的梯度设置为 0，因此 $\boldsymbol{W}$ 必须与 $\boldsymbol{\theta}$ 成比例。借鉴 UFO-MKL 的证明[239], 显式地给出 $\boldsymbol{W}$ 与 $\boldsymbol{\theta}$ 的关联

$$\boldsymbol{W}_m = \mu_m \boldsymbol{\theta}_m$$

基于上述关联，算法 2 通过 $\nabla\Omega^*(\boldsymbol{\theta})$ 同时对 $\boldsymbol{W}, \boldsymbol{\mu}$ 进行了更新。为方便推导，主要考虑 $c_m = \mu_m \|\boldsymbol{\theta}_m\|$，优化问题 (5.4) 重写为:

$$\arg\min_{\boldsymbol{c}} (\beta \mathbf{r} - \mathbf{a}) \cdot \boldsymbol{c} + \frac{\alpha}{2} \|\boldsymbol{c}\|_p^2 \tag{5.5}$$

其中 $\mathbf{a} = \Big[\|\boldsymbol{\theta}_1\|, \ldots, \|\boldsymbol{\theta}_M\|\Big]$。

上述优化问题的最优条件说明 $\boldsymbol{c}^*$ 是 (5.5) 的最优解[243]。通过将上述最小化问题 (5.5) 的梯度设置为 0，可得

$$\beta \mathbf{r} - \mathbf{a} + \alpha \boldsymbol{c}^* = 0. \tag{5.6}$$

进而可以得到 $\boldsymbol{c}^* = \frac{1}{\alpha}\Big(\mathbf{a} - \beta \mathbf{r}\Big)$。使用引理 5.2、引理 5.3，可以得到如下闭式解

$$c_m = f^{-1}(c_m^*) = \nabla_m \left(\frac{1}{2}\|c_m^*\|_q^2\right) = \frac{\operatorname{sgn}(c_m^*)|c_m^*|^{q-1}}{\alpha \|\boldsymbol{c}^*\|_q^{q-2}}.$$

同时令

$$\boldsymbol{\nu} = \Big[\|\boldsymbol{\theta}_1\| - \beta \mathbf{r}_1, \ldots, \|\boldsymbol{\theta}_M\| - \beta \mathbf{r}_M\Big],$$

以及 $\mu_m = c_m / \|\boldsymbol{\theta}_m\|$、$\boldsymbol{W}_m = \mu_m \boldsymbol{\theta}_m$，可得多核组合系数为

$$\mu_m = \frac{\operatorname{sgn}(\nu_m)|\nu_m|^{q-1}}{\alpha \|\boldsymbol{\theta}_m\| \|\boldsymbol{\nu}\|_q^{q-2}},$$

原始权重为

$$\boldsymbol{W}_m = \frac{\operatorname{sgn}(\nu_m)\boldsymbol{\theta}_m|\nu_m|^{q-1}}{\alpha \|\boldsymbol{\theta}_m\| \|\boldsymbol{\nu}\|_q^{q-2}},$$

其中 $\operatorname{sgn} : \mathbb{R} \to \mathbb{R}$ 为符号函数，在求解次微分的过程中产生。 Xiao 在其研究工作[242] 的第 7.2 章针对 ℓ-1 正则化的最优化问题给出了相似的分析。 □

5.4.2 近端梯度下降

5.4.2.1 构造优化目标

使用随机特征 $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$ 对核函数进行加速 $\kappa(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \phi_M(\boldsymbol{x}), \phi_M(\boldsymbol{x}') \rangle$，并在特征空间中使用线性形式得到预测模型 $f(\boldsymbol{x}) = \boldsymbol{W}^\top \phi_M(\boldsymbol{x})$。

将 $\boldsymbol{S}$ 定义为全部数据 $D^l \cup D^u$ 上的相似度矩阵，其中 $\boldsymbol{S}_{ij}$ 代表输入 $\mathbf{x}_i$、输入 $\mathbf{x}_j$ 的相似性，比如可以使用热核函数进行独立 $S_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$。基于半监督学习的连续性假设：相似输入样本对应的标签也是相似的，将半监督近似核方法的代价函数定义为

$$E(f) = \sum_{i,j=1}^{n+u} S_{ij}\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 = \mathrm{trace}(\boldsymbol{W}^\top \phi_M(\mathbf{X})\mathbf{L}\phi_M(\mathbf{X})^\top \mathbf{W}),$$

其中 $\phi_M(\mathbf{X}) \in \mathbb{R}^{M\times(n+u)}$ 为全部样本输入对应的随机特征。图 Laplacian 矩阵定义为 $\mathbf{L} = \mathbf{D} - \boldsymbol{S}$，其中 $\mathbf{D}$ 为对角矩阵，对角元素为 $\mathbf{D}_{ii} = \sum_{j=1}^{n+u} \boldsymbol{S}_{ij}$。为最小化代价函数 $E(f)$，将其加入到经验风险最小化优化目标中。

根据泛化理论分析中定理 3.12，线性学习器的局部 Rademacher 复杂度由权重矩阵的尾部奇异值之和 $\sum_{j>\theta} \lambda_j(\boldsymbol{W})$ 界定，为获得更好的泛化性能将最小化 $\boldsymbol{W}$ 的尾部奇异值之和加入到经验风险最小化学习 (ERM) 框架中。结合了 Laplacian 正则化（用于利用无标签数据）、尾部奇异值之和（用于界定局部 Rademacher 复杂度）的近似核方法的优化目标为

$$\arg\min_{f\in H_\kappa} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i) + \lambda_I \mathrm{trace}(\boldsymbol{W}^\top \phi_M(\mathbf{X})\mathbf{L}\phi_M(\mathbf{X})^\top \boldsymbol{W})}_{g(\boldsymbol{W})} + \lambda_S \sum_{j>\theta} \lambda_j(\boldsymbol{W}). \quad (5.7)$$

将目标函数 (5.7) 划分为对于 $\boldsymbol{W}$ 可微分、不可微分两部分。 $g(\boldsymbol{W})$ 包括经验风险、Laplacian 正则化项，对于 $\boldsymbol{W}$ 是可微的。而 $\boldsymbol{W}$ 的尾部奇异值之和 $\sum_{j>\theta} \lambda_j(\boldsymbol{W})$ 不可微。已完成工作[207,208] 针对不可微的优化目标函数，并使用近端梯度下降的奇异值阈值方法 (singular value thresholding, SVT) 求解优化目标。

5.4.2.2　使用近端梯度下降算法求解不可微优化目标

SVT 算法最早由 Cai 等提出[244]，用于求解最小化矩阵迹问题（全部奇异值之和）； Xu 等为求解最小化尾部奇异值之和的问题，发展出 PSVT 算法[215]；Lu 等为给出更广义的 GSVT 算法，能够求解最小化部分奇异值之和[245]。本文借鉴 PSVT 算法、GSVT 算法思想，对 $\boldsymbol{W}$ 更新两次：首先针对可微部分 $g(\boldsymbol{W})$ 使用随机梯度下降算法更新 $\boldsymbol{W}$；之后更新 $\boldsymbol{W}$ 较大的奇异值，从而更新 $\boldsymbol{W}$。

算法 3 近端梯度下降算法

输入 有标签数据集 D^l、无标签数据集 D^u。初始化矩阵 $\boldsymbol{W}_1 = \mathbf{0}$，总迭代次数 T，特征映射函数 $\phi_M : \mathbb{R}^d \to \mathbb{R}^M$。初始化超参数 $\theta, \lambda_I, \lambda_S$、步长 η_t。

输出 $\boldsymbol{W}_{T+1}$

在全部数据 D^l、D^u 上计算 Laplacian 矩阵 $\boldsymbol{L}$。

在全部数据上使用特征映射：$\widetilde{\mathbf{X}} = \phi_M(\mathbf{X})$。

for $t = 1, 2, \cdots, T$ **do**

在有标签数据集上随机采样 b 个样本 $(\boldsymbol{x}_i,\ \boldsymbol{y}_i)_{i=1}^b \in D^l$ 作为一个 mini-batch。

在 mini-batch 上计算可微部分 $\nabla g(\boldsymbol{W}_t)$ 梯度

$$\nabla g(\boldsymbol{W}_t) = \frac{1}{b}\sum_{i=1}^{b} \frac{\partial\ \ell(h(\boldsymbol{x}_i),\ \boldsymbol{y}_i)}{\partial\ \boldsymbol{W}_t} + 2\lambda_I \widetilde{\mathbf{X}}\boldsymbol{L}\widetilde{\mathbf{X}}^\top W_t. \tag{5.8}$$

梯度下降更新 $\boldsymbol{W}$，并进行奇异值分解

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{W}^t - \eta_t \nabla g(\boldsymbol{W}^t) \tag{5.9}$$

更新 $\boldsymbol{W}$ 奇异值，从而更新 $\boldsymbol{W}$

$$\boldsymbol{W}^{t+1} = \boldsymbol{U}\boldsymbol{\Sigma}_\tau^\theta \boldsymbol{V}^\top \ \text{ 其中 } \ \tau = \eta_t \lambda_S \tag{5.10}$$

end for

在每次迭代中, 为获得等式 (4.10) 紧的代理函数，保持 $\|\boldsymbol{W}\|_*$ 不变而只对可微分部分 $g(\boldsymbol{W})$ 进行放缩，从而应用近端梯度下降算法[246]

$$\begin{aligned}
\boldsymbol{W}^{t+1} &= \arg\min_{\boldsymbol{W}} \lambda_S \sum_{j>\theta} \lambda_j(\boldsymbol{W}) + g(\boldsymbol{W}) \\
&= \arg\min_{\boldsymbol{W}} \lambda_S \sum_{j>\theta} \lambda_j(\boldsymbol{W}) + g(\boldsymbol{W}^t) + \langle \nabla g(\boldsymbol{W}^t), \boldsymbol{W} - \boldsymbol{W}^t \rangle + \frac{1}{2\eta}\|\boldsymbol{W} - \boldsymbol{W}^t\|_F^2 \\
&= \arg\min_{\boldsymbol{W}} \lambda_S \sum_{j>\theta} \lambda_j(\boldsymbol{W}) + \frac{1}{2\eta}\|\boldsymbol{W} - (\boldsymbol{W}^t - \eta \nabla g(\boldsymbol{W}^t)\|_F^2 \\
&= \arg\min_{\boldsymbol{W}} \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{Q}\|_F^2 + \eta\lambda_S \sum_{j>\theta} \lambda_j(\boldsymbol{W}),
\end{aligned} \tag{5.11}$$

其中 $\boldsymbol{Q} = \boldsymbol{W}^t - \eta \nabla g(\boldsymbol{W}^t)$ 为中间变量、 η 为学习率。

引理 5.4 (定理 6[215]). 令 $\boldsymbol{Q} \in \mathbb{R}^{M\times K}$ 的秩为 r，它的奇异值分解为 $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$，其中 $\boldsymbol{U} \in \mathbb{R}^{d\times r}$，$\boldsymbol{V} \in \mathbb{R}^{K\times r}$ 中列是相互正交的。$\boldsymbol{\Sigma}$ 为对角矩阵，对角元素为 $diag(\{\sigma_i\}_{1\le i\le r})$。存在如下闭式解

$$\arg\min_{\boldsymbol{W}}\left\{\frac{1}{2}\|\boldsymbol{W}-\boldsymbol{Q}\|_F^2+\tau\sum_{j>\theta}\lambda_j(\boldsymbol{W})\right\}=\boldsymbol{U}\boldsymbol{\Sigma}_\tau^\theta\boldsymbol{V}^{\top},$$

其中，对角矩阵只有前 θ 大的奇异值进行了更新

$$(\boldsymbol{\Sigma}_\tau^\theta)_{jj}=\begin{cases}\max(0,\boldsymbol{\Sigma}_{jj}-\tau), & i\le\theta,\\ \boldsymbol{\Sigma}_{jj}, & i>\theta.\end{cases}$$

将等式 (5.11)、引理 5.4 相结合，应用到优化目标 (5.7)，得到最终求解算法（算法 3）。如算法 3 所示，近端梯度下降算法更新 $\boldsymbol{W}$ 两次：

- 第一次更新 (5.9)：可微部分 $g(\boldsymbol{W})$ 使用随机梯度下降更新 $\boldsymbol{W}$。
- 第二次更新 (5.10)：更新 $\boldsymbol{W}$ 部分奇异值，从而更新 $\boldsymbol{W}$。

由于未具体定义损失函数，所以等式 (5.8) 对于损失的微分 $\frac{\partial\,\ell(h(\boldsymbol{x}_i),\,\boldsymbol{y}_i)}{\partial\,\boldsymbol{W}_t}$ 是未知的。下面介绍分别介绍多分类、多标签两个例子，确定具体的损失函数，从而给出完整的学习过程。

5.4.2.3　求解多分类问题

考虑类别数为 K 的多分类问题. 输出空间使用独热形式 $\mathcal{Y}=\{0,1\}^K$，即只有对应类别下标的元素为 1，其他 $K-1$ 个元素均为 0。对于标签

$$\boldsymbol{y}_i=[0,\cdots,0,1,0,\cdots,0]^{\top},$$

只有对应类别位置的元素标注为 1。将多分类的边界 (margin) 定义为

$$m_f(\boldsymbol{x}_i,\,\boldsymbol{y}_i)=[f(\boldsymbol{x}_i)]^{\top}\boldsymbol{y}_i-\max_{\boldsymbol{y}_i'\ne\boldsymbol{y}_i}[f(\boldsymbol{x}_i)]^{\top}\boldsymbol{y}_i'.$$

如果边界 $m_f(\boldsymbol{x}_i,\,\boldsymbol{y}_i)\le 0$，则学习器 f 误分类了样本 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$。若损失函数为 0-1 损失，则有 $\ell(f(\boldsymbol{x}_i),\,\boldsymbol{y}_i)=1_{m_f(\boldsymbol{x}_i,\,\boldsymbol{y}_i)\le 0}$。由于 0-1 不连续，很难优化，因此考虑使用连续的、能够界定 0–1 损失的损失函数。对于多分类问题使用合页损失

$$\ell(f(\boldsymbol{x}_i),\,\boldsymbol{y}_i)=|1-m_f(\boldsymbol{x}_i,\,\boldsymbol{y}_i)|_+.$$

由于合页损失在 $m_f(\boldsymbol{x}_i,\ \boldsymbol{y}_i) = 0$ 的情况下是不可微的，因此使用求解次微分代替求解梯度。最终，多分类对应经验损失的次微分可以定义为

$$\frac{\partial\ \ell(f(\boldsymbol{x}_i),\ \boldsymbol{y}_i)}{\partial\ \boldsymbol{W}_t} = \begin{cases} \boldsymbol{0}_{M\times K}, & m_f(\boldsymbol{x}_i,\ \boldsymbol{y}_i) \geq 1, \\ \phi_M(\boldsymbol{x})[\boldsymbol{y}_i' - \boldsymbol{y}_i]^\top, & \text{otherwise}, \end{cases} \tag{5.12}$$

其中 $(\boldsymbol{x}_i,\ \boldsymbol{y}_i)$ 为有标签数据集 D^l 中的某个样本。将多分类次微分 (5.12) 带入到算法 3，即可求得对应多分类问题的学习器。

5.4.2.4 求解多标签问题

多标签问题包括两种：多标签分类 $\mathcal{Y} = \{0, 1\}^K$ 、多标签回归 $\mathcal{Y} = \mathbb{R}^K$。对于多标签问题，均使用平方损失

$$\ell(f(\boldsymbol{x}_i),\ \boldsymbol{y}_i) = \|\boldsymbol{y} - f(\boldsymbol{x}_i)\|_2^2.$$

多标签经验误差的梯度为

$$\frac{\partial\ \ell(f(\boldsymbol{x}_i),\ \boldsymbol{y}_i)}{\partial\ \boldsymbol{W}_t} = 2\phi_M(\boldsymbol{x}_i)[f(\boldsymbol{x}_i) - \boldsymbol{y}_i]^\top, \tag{5.13}$$

其中 $(\boldsymbol{x}_i,\ \boldsymbol{y}_i)$ 为有标签数据集 D^l 中的某个样本。将多标签次微分 (5.13) 带入到算法 3，即可求得对应多标签问题的学习器。

第6章　实验分析

首先介绍实验中使用的评价指标、数据来源、超参选择、测试结果稳定性：

1. 评价指标

令 $D^t = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n_t}$ 为测试数据，将最终选取（学习）的核函数对应模型的泛化性能作为评价指标。使用到的评价指标包括：

- 分类问题：分类错误率 (Error) 作为评价指标

$$\text{Error} = \frac{1}{n_t K} \sum_{i=1}^{n_t} \sum_{k=1}^{K} y'_{ik} \oplus y_{ik}.$$

其中，$\boldsymbol{y}'_i$ 为预测标签、$\boldsymbol{y}_i$ 为真实标签、$\oplus$ 为异或运算。分类问题的准确率指标为 $\text{Accuracy} = 1 - \text{Error}$。

- 回归问题：均方根误差 (root mean square error, RMSE) 作为评价指标

$$\text{RMSE} = \frac{1}{n_t K} \sum_{i=1}^{n_t} \|f(\boldsymbol{x}_i) - \boldsymbol{y}_i\|_2.$$

2. 实验数据来源

实验数据取自 UCI [1] 和 LIBSVM [2]。由于面向学习任务不同，所以在实验中没有使用完全相同的数据集。

3. 超参数选择

实验中用到的多个超参数，并使用 10–折交叉验证选取在候选参数集合中选取平均验证误差最小的超参数组合。用到的超参数包括正则化系数 $\lambda_A, \lambda_I, \lambda_S \in \{10^{-15}, 10^{-11}, \cdots, 10^{-1}\}$、核超参数比如高斯核 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp(\frac{-\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2\sigma^2})$ 中 $\sigma \in \{2^{-15}, 2^{-14}, \cdots, 2^{15}\}$、矩阵截断点 $\theta \in \{0, 0.1, \cdots, 0.9\} \times n$、构造 Laplacian 矩阵时用到 k–近邻算中的 $k \in \{2, 4, 6\}$。

4. 测试结果稳定性

为获得稳定的测试结果，在重复划分训练集、测试集 T 次 ($T \geq 10$)，并在每一次划分上运行所有算法，从而每种算法在每个数据集上获得 T 个测试结果。基于每种算法的 T 个测试结果，使用 t 检验比较对对比算法进行比较，以 95%

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

的显著性水平统计方法间显著性差异。表格中加粗最优的测试结果，并使用下划线标注与最优测试结果无显著差异的测试结果。

6.1 大规模半监督核方法模型选择准则

本节使用实验验证第 4 章提到的核方法模型选择准则，包括：最大化谱度量、最小化核矩阵尾部特征值之和、反向传播更新核超参数。这三种核方法模型选择准则适用问题、应用方式不尽相同，所以实验数据集、评价指标不同。

6.1.1 最大化谱度量

最大化谱度量适用于二分类有监督问题。在核函数候选集 $\mathcal{K}$ 中选取使得谱度量最大的核函数作为准则

$$\arg\max_{\kappa\in\mathcal{K}} \mathrm{SM}(\kappa,\psi) = \frac{1}{n}\bar{\boldsymbol{y}}^{\mathrm{T}}\mathbf{N}^{r}\bar{\boldsymbol{y}}.$$

6.1.1.1 平均测试误差、训练时间

首先将最大化谱度量与其他 5 个常用的核方法模型选择准则进行对比:

(1) 5–折交叉验证 (5-folds cross-validation, CV)

(2) 高效留一法 (efficient leave-one-out cross-validation, ELOO)[247]

(3) 中心化核对齐 (centered kernel target alignment, CKTA)[230]

(4) 基于特征空间的核矩阵评价 (feature space-based kernel matrix evaluation, FSM)[248]

(5) 特征值比率 (eigenvalue ratio, ER)[231]

对于每个数据集，基于所有不同的核选择准则，在训练集上选取高斯核超参数 σ。使用选取的核函数，并重复划分训练集、测试集，记录重复的测试结果。表 6.1 报告了平均测试误差、标准差，从该表中可以得出

(1) 在几乎所有的数据集上，SM 显著优于 FSM 和 CKTA。这是因为 FSM 和 CKTA 对学习算法没有理论保证，FSM、CKTA 方法所选择的内核并不能保证良好的泛化性能。

(2) 相比于 CV 方法，SM 在 25 个数据集中的 19 个数据集上测试误差更低；相比于 ER 方法， SM 在 25 个数据集中的 17 个数据集上测试误差更低。因此，在大多数数据集上 SM 优于 CV 和 ER。

表 6.1 最大化谱度量相关方法的平均分类错误率 (%) 对比

	SM	CV	ELOO	CKTA	FSM	ER
a1a	**16.84±1.39**	17.02±1.57	16.88±1.41	18.86±1.49	24.72±1.67	16.97±1.52
a2a	**17.78±1.28**	17.96±1.25	17.94±1.27	18.52±1.26	25.62±1.47	18.99±1.37
anneal	**2.69±3.28**	3.81±4.11	**2.69±3.28**	4.75±4.78	5.13±4.18	5.50±4.95
australian	13.71±2.10	13.84±2.18	13.82±2.04	13.91±1.89	44.71±2.47	**13.53±2.06**
autos	**11.81±11.67**	**11.81±11.67**	12.75±11.06	13.71±12.03	12.71±8.06	12.14±11.51
breast-w	**3.27±1.01**	3.56±1.16	3.59±1.08	3.51±1.05	3.50±1.05	4.26±1.40
breast-cancer	**3.18±1.15**	3.63±1.16	3.50±1.23	3.63±1.16	3.60±1.14	4.04±1.12
bupa	30.29±3.48	**29.10±4.04**	30.31±4.27	35.81±3.45	39.77±3.68	29.13±4.46
colic	**15.62±3.00**	16.47±2.78	15.73±2.97	19.27±2.58	36.42±3.28	17.35±3.09
diabetes	24.22±2.41	24.69±2.71	**23.51±2.75**	24.85±2.46	35.30±3.00	23.90±2.48
glass	22.09±5.07	21.82±5.68	**20.95±4.82**	26.41±7.13	43.00±9.22	22.50±5.08
german.numer	24.09±2.15	25.28±2.38	**23.81±2.26**	26.02±2.16	29.89±2.41	25.33±2.14
heart	16.53±3.27	16.69±3.36	**15.95±3.29**	18.67±3.78	44.37±5.50	15.98±3.47
hepatitis	**15.57±4.68**	17.09±5.74	16.63±4.64	15.74±5.00	21.22±5.41	18.91±6.20
ionosphere	4.88±2.10	5.28±2.11	6.42±2.17	11.70±3.43	35.77±4.00	**4.86±1.99**
labor	**13.65±8.10**	14.47±8.08	14.82±8.34	15.41±8.80	34.59±8.70	18.82±8.81
pima	23.80±2.14	22.78±2.36	**22.51±2.41**	24.38±2.28	34.47±2.42	22.78±2.07
segment	**0.01±0.00**	0.06±0.24	0.20±0.04	0.32±0.03	0.21±0.01	0.24±0.04
liver-disorders	31.94±3.21	**29.00±4.11**	30.02±4.76	36.27±3.93	40.90±4.10	29.69±4.97
sonar	15.06±4.80	14.26±4.93	**13.68±4.43**	15.00±5.51	49.32±6.93	18.84±5.75
vehicle	**3.02±1.79**	3.33±1.77	**3.02±1.79**	3.77±1.51	53.32±3.38	5.52±2.44
vote	**4.31±1.71**	4.78±1.74	4.82±1.73	5.25±1.72	6.37±3.96	7.80±2.33
wpbc	23.10±4.58	22.83±4.32	21.93±4.45	**21.87±4.13**	22.13±4.19	**21.87±4.13**
tic-tac-toe	10.10±1.93	10.28±1.66	**9.78±1.66**	33.62±5.31	34.44±2.04	14.62±2.05
wdbc	**2.29±1.15**	2.43±1.07	2.73±1.11	2.82±1.20	37.49±3.83	4.75±1.66

(3) SM 给出了与 ELOO 相似的测试结果。在 9 个数据集上，SM 显著优于 ELOO (autos、breast-w、breast-cancer、hepatitis、ionosphere、labor、segment、vote and wdbc)；而在 8 个数据集上，SM 显著差于 ELOO (diabetes、glass、heart、pima、liver-disorders、sonar、wpbc and tic-tac-toe)。

表 6.2 记录对比方法的训练时间。结果显示 SM 比 CV、ELOO、ER 三种方法训练时间都短，并与 CKTA、FSM 训练时间相似。实际训练时间表现与表 4.1 中的时间复杂度结果一致。综合表 6.1 中测试误差对比与表 6.2 中训练时间对比，可以看出 SM 可以保证良好的泛化性能同时具有较高的计算效率。

表 6.2 最大化谱度量相关方法的训练时间（秒）对比

	SM	CV	ELOO	CKTA	FSM	ER
a1a	**63.77**	704.50	217.03	87.51	79.71	422.71
a2a	**124.53**	1995.76	810.87	172.28	156.68	1558.70
anneal	**0.28**	4.60	0.84	0.43	0.55	1.36
australian	**8.30**	111.31	30.33	11.34	9.51	52.53
autos	**0.10**	2.18	0.23	0.17	0.29	0.37
breast-cancer	**8.05**	104.55	27.13	11.10	9.36	47.68
breast-w	**8.59**	105.39	29.77	11.80	10.11	49.86
colic	**1.55**	25.72	6.52	2.15	2.17	11.60
glass	**0.33**	5.74	1.15	0.53	0.66	2.00
heart	**1.02**	14.71	3.66	1.12	1.20	6.33
hepatitis	**0.38**	6.55	1.42	0.58	0.70	2.44
ionosphere	**1.59**	24.60	6.08	2.03	2.06	10.60
labor	**0.16**	2.88	0.42	0.28	0.42	0.67
pima	**10.99**	137.19	36.60	15.11	12.74	62.45
segment	**7.50**	91.73	23.83	10.49	8.93	42.63
diabetes	**10.70**	134.71	36.26	14.90	12.51	62.46
german.numer	**21.80**	249.98	72.63	29.59	26.27	127.51
liver-disorders	**1.37**	21.04	5.46	1.91	1.94	9.27
sonar	**0.62**	10.53	2.34	0.87	0.98	4.06
vehicle	**2.03**	35.64	9.35	2.86	2.85	15.95
vote	**2.03**	34.43	9.33	2.79	2.81	15.92
wpbc	**0.52**	9.06	2.07	0.76	0.92	3.33
bupa	**1.36**	21.09	5.34	1.85	1.88	9.19
tic-tac-toe	**18.93**	224.13	63.10	26.13	23.04	107.12
wdbc	**5.75**	61.32	19.18	7.86	6.80	33.65

6.1.1.2 谱度量中参数 r 的影响

图 6.1 给出不同 r 取值下的平均测试误差。对于任意固定的 r，在训练集上使用最大化谱度量选取高斯核超参数 σ，并对选出的核模型进行训练、测试。实验发现对于大多数数据集 $r \in [2, 5]$ 比较合适。同时，经过多次实验发现，将 r 随机设置为 $r \in \{2, 3, 4\}$ 不会带来显著的精度损失。

6.1.2 最小化核矩阵尾部特征值之和

最小化核矩阵尾部特征值之和适用于多核学习问题，实验中使用多核多分类验证准则有效性。将核矩阵尾部特征值放在经验误差最小化 (ERM) 学习框架

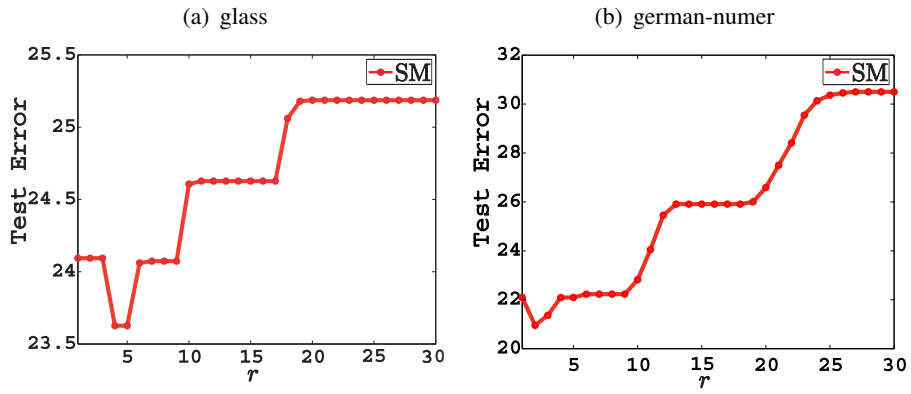

图 6.1 谱度量不同 r 对测试误差的影响

中，同时学习多核组合系数 $\boldsymbol{\mu}$、核模型权重 $\boldsymbol{W}$。优化目标 (4.4) 为

$$\min_{\boldsymbol{W},\boldsymbol{\mu}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i) + \frac{\alpha}{2}\|\boldsymbol{W}\|_{2,p}^2 + \beta \sum_{m=1}^{M} \mu_m r_m,$$

本节面向多分类学习，将提出的 Conv-MKL 算法（算法 1）、SMSD-MKL 算法（算法 2），与其他 7 种主流多分类学习算法进行对比:

- 一对一方法 (One-against-One)[249]。
- 一对多方法 (One-against-the-Rest)[250]。
- ℓ_1 正则化的线性多分类 (LMC)[251]。
- 极小范数学习器 (generalized minimal norm problem solver, GMNP)[252]。
- ℓ_1 范数的多核多分类 (ℓ_1 MC-MKL)[253]。
- ℓ_2 范数的多核多分类 (ℓ_2 MC-MKL)[253]。
- 混合范数的多核学习器 (UFO-MKL)[239]。

表 6.3 记录了所有对比方法的平均分类准确率、标准差，其中结果说明

(1) 与传统的多核学习方法相比，除了 UFO–MKL 在数据集 *satige* 结果外，使用最小化多核矩阵尾部特征值之和的 Conv-MKL、SMSD-MKL 在其他所有数据集上给出更高的准确率。因此，Conv-MKL、SMSD-MKL 优于其他多核学习方法。

(2) SMSD-MKL 在 2/3 比例的数据集上表现比 Conv-MKL 更好，因此 SMSD-MKL 泛化性能更稳定。

(3) 多核学习方法的测试结果全部优于单核 (One vs. One、One vs. Rest、GMNP) 测试结果。

(4) 在所有数据集上，任意核化的多分类学习器都要比线性多分类 (LMC) 分类准确要高。

表 6.3 最小化核矩阵尾部特征值之和相关算法的分类准确率 (%) 对比

	Conv-MKL	SMSD-MKL	LMC	One vs. One	One vs. Rest	GMNP	ℓ_1 MC-MKL	ℓ_2 MC-MKL	UFO-MKL
plant	77.14±2.25	**78.01±2.17**	70.12±2.96	75.83±2.69	75.17±2.68	75.42±3.64	77.60±2.63	75.49±2.48	76.77±2.42
psortPos	74.41±3.35	**76.23±3.39**	63.85±3.94	73.33±4.21	71.70±4.89	73.55±4.22	71.87±4.87	70.70±4.89	74.56±4.04
psortNeg	74.07±2.16	**74.66±1.90**	57.85±2.49	73.74±2.87	71.94±2.50	74.27±2.51	72.83±2.20	72.42±2.65	73.80±2.26
nonpl	**79.15±1.51**	78.69±1.58	75.16±1.48	77.78±1.52	77.49±1.53	78.35±1.46	77.89±1.79	77.95±1.64	78.07±1.56
sector	92.83±2.62	**93.39±0.70**	93.16±0.66	90.61±0.69	91.34±0.61	\	\	92.15±2.57	92.60±0.47
segment	96.79±0.91	**97.62±0.83**	95.07±1.11	97.08±0.61	97.02±0.80	96.87±0.80	96.98±0.64	97.58±0.68	97.20±0.82
vehicle	**79.35±2.27**	77.28±2.78	75.61±3.56	78.72±1.92	79.11±1.94	81.57±2.24	74.96±2.93	76.27±3.15	76.92±2.83
vowel	98.82±1.19	**98.83±5.57**	62.32±4.97	98.12±1.76	98.22±1.83	97.04±1.85	98.27±1.22	97.86±1.75	98.22±1.62
wine	**99.63±0.96**	**99.63±0.96**	97.87±2.80	97.24±3.05	98.14±3.04	97.69±2.43	98.61±1.75	98.52±1.89	99.44±1.13
dna	96.08±0.83	**96.30±0.79**	92.02±1.50	95.89±0.56	95.61±0.73	94.60±0.94	96.27±0.68	95.06±0.92	95.84±0.61
glass	**75.19±5.05**	73.72±5.80	63.95±6.04	71.98±5.75	70.00±5.75	71.24±8.14	69.07±8.08	74.03±6.41	72.46±6.12
iris	96.67±2.94	**97.00±2.63**	88.00±7.82	95.93±3.25	95.87±3.20	95.40±7.34	95.40±6.46	94.00±7.82	95.93±2.88
svmguide2	82.69±5.65	**85.17±3.83**	81.10±4.15	84.79±3.45	84.27±3.03	81.77±3.45	83.16±3.63	83.84±4.21	82.91±3.09
satimage	91.64±0.88	91.78±0.82	84.95±1.15	90.67±0.91	89.29±0.96	89.97±0.81	91.86±0.62	90.43±1.27	**91.92±0.83**

上述结果表明，使用最小化核矩阵尾部特征值（界定最小化局部 Rademacher 复杂度）可以显著地提高多类多核学习算法的性能，与泛化理论分析（推论 3.11）是一致的。

6.1.3 反向传播更新核超参数

反向传播更新核超参数适用于谱核方法，实验中使用多分类、回归问题对模型选择准则进行验证。以端到端形式，根据优化目标，同时更新谱核密度对应频率矩阵 Ω、Ω'，以及核模型权重 $\boldsymbol{W}$。优化目标 (4.10) 可以写为

$$\arg\min_{\boldsymbol{W},\Omega,\Omega'} \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i),\boldsymbol{y}_i) + \lambda_A\|\boldsymbol{W}\|_* + \lambda_B\|\phi_M(\boldsymbol{X})\|_F^2.$$

为验证学习框架的有效性，将图 4.1 自动谱核学习 (ASKL) 框架，与其他几种谱核学习方法进行对比，在表 6.4 中列出，均为 ASKL 的特例：

- **SK**：随机傅里叶特征近似平稳谱核[31]。
- **NSK**：随机傅里叶特征近似非平稳谱核[101]。
- **SKL**：使用反向传播学习的平稳谱核[254]。
- **NSKL**: 使用反向传播学习的非平稳谱核。 NSKL 是 ASKL 的特例，仅使用 $\|\boldsymbol{W}\|_F^2$ 作为正则化项。

表 6.4 自动谱核学习 (ASKL) 对比算法

方法	核函数	核超参数	正则化项
SK	平稳谱核	交叉验证选择	$\|\boldsymbol{W}\|_F^2$
NSK	非平稳谱核	交叉验证选择	$\|\boldsymbol{W}\|_F^2$
SKL	平稳谱核	反向传播更新	$\|\boldsymbol{W}\|_F^2$
NSKL	非平稳谱核	反向传播更新	$\|\boldsymbol{W}\|_F^2$
ASKL	非平稳谱核	反向传播更新	$\|\boldsymbol{W}\|_*, \|\phi(\boldsymbol{X})\|_F^2$

6.1.3.1　泛化性能对比

在表 6.5 中，对于分类数据集使用准确率 (%) 作为评价指标，(↑) 代表准确率越高泛化性能越好；对于回归数据集使用归一化的均方根误差 (RMSE) 作为评价指标，(↓) 代表 RMSE 越低泛化性能越好。表 6.5 中结果说明:

(1) 使用反向传播的自动谱核学习算法 ASKL 在所有数据集上取得了最优结果，验证了泛化理论分析的结果（推论 4.3）。

(2) 使用非平稳谱核带来了显著提升，但平稳谱核在相对简单的数据集上也表现良好，比如 *shuttle*、*cpusmall*。

(3) 核超参数是否使用反向传播更新，为泛化性能带来了巨大差异。未使用反向传播的 {SK, NSK} 比使用反向传播的 {SKL, NSKL} 测试结果表现差距显著，尤其是在复杂度的数据集上，比如 *satimage*、*letter*。这说明了反向传播更新核超参数能够带来泛化性能显著的提升。

(4) 自动谱核学习 ASKL 的测试性能比 NSKL 更优，说明两个正则化项 $\|\boldsymbol{W}\|_*$、$\|\phi(\boldsymbol{X})\|_F^2$ 提升了谱核学习的泛化性能。

实验结果表明：通过使用非平稳谱核、反向传播更新核超参数、正则化项 $\|\boldsymbol{W}\|_*$ 及 $\|\phi(\boldsymbol{X})\|_F^2$，自动谱核学习 ASKL 获得了良好的泛化性能，与推论 4.3 中理论结果一致。

6.1.3.2　收敛情况对比

在手写字符识别数据集 (MNIST) 上记录自动谱核学习 ASKL 及其对比算法的收敛情况。在迭代过程中，每 200 次迭代记录的分类精度和目标函数取值，批大小为 32。精度曲线与目标函数曲线是相关的，目标函数越小，精度越高。图

表 6.5 自动谱核学习 (ASKL) 相关对比算法的平均测试结果对比

		SK	NSK	SKL	NSKL	ASKL
Accuracy(↑)	segment	89.93±2.12	90.15±2.08	94.58±1.86	94.37±0.81	**95.02±1.54**
	satimage	74.54±1.35	75.15±1.38	83.61±1.08	83.74±1.34	**85.32±1.45**
	USPS	93.19±2.84	93.81±2.13	95.13±0.91	95.27±1.65	**97.76±1.14**
	pendigits	96.93±1.53	97.39±1.41	98.19±2.30	98.28±1.68	**99.06±1.26**
	letter	76.50±1.21	78.21±1.56	93.60±1.14	94.66±2.21	**95.70±1.74**
	porker	49.80±2.11	51.85±0.97	54.27±2.72	54.69±1.68	**54.85±1.28**
	shuttle	98.17±2.81	98.21±1.46	98.87±1.42	98.74±1.07	**98.98±0.94**
	MNIST	96.03±2.21	96.45±2.16	96.67±1.61	98.03±1.16	**98.26±1.78**
RMSE(↓)	abalone	10.09±0.42	9.71±0.28	8.35±0.28	**7.85±0.42**	7.88±0.16
	space_ga	11.86±0.26	11.58±0.42	11.40±0.18	11.39±0.46	**11.34±0.27**
	cpusmall	2.77±0.71	2.84±0.38	2.56±0.72	2.57±0.63	**2.42±0.48**
	cadata	50.31±0.92	51.47±0.32	47.67±0.33	47.71±0.30	**46.34±0.23**

6.2 说明自动谱核学习 ASKL 以更快的收敛速度达到了更小的经验误差。

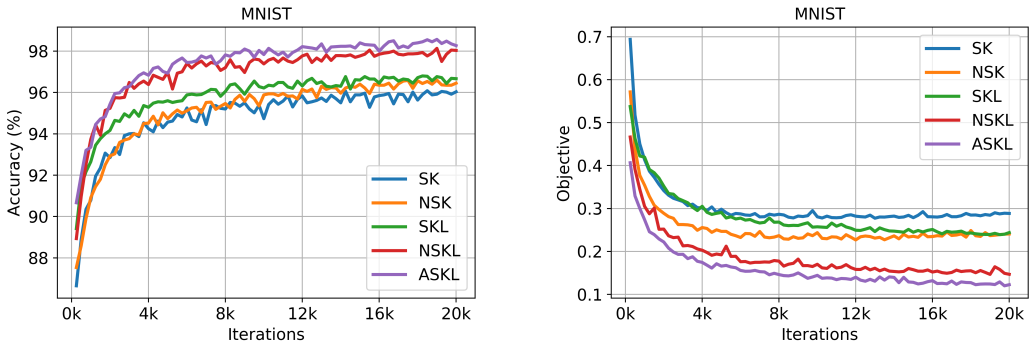

图 6.2 左图 MNIST 上的准确率曲线、右图 MNIST 上的目标函数曲线

6.2 大规模半监督核方法模型选择算法

由于核方法的时间、空间复杂度过高，均不低于 $\mathcal{O}(n^2)$，单独使用一种加速方法（如分布式、Nyström 采样、随机特征、一阶梯度随机优化）后仍不能适用于大规模数据，因此往往将多种加速手段相结合来训练大规模核学习器。

本节对第 5 章中介绍的核方法加速手段进行组合，并通过实验验证多种加

表 6.6 Nyström-PCG 相关算法平均测试误差对比

数据集	样本维度	RLS-CG	LapRLS-CG	LapRLS-PCG	Nyström-CG	Nyström-PCG
madelon	2000	1.036±0.009	**0.990±0.007**	**0.990±0.007**	0.991±0.009	0.991±0.009
space_ga	3107	1.251±0.004	**1.210±0.004**	**1.210±0.004**	**1.210±0.004**	**1.210±0.004**
abalone	4177	$4.55\pm0.2\times10^{3}$	$\mathbf{4.17\pm0.1\times10^{3}}$	$\mathbf{4.17\pm0.1\times10^{3}}$	$4.18\pm0.1\times10^{3}$	$4.18\pm0.1\times10^{3}$
phishing	11055	0.426±0.049	0.294±0.005	**0.273±0.007**	0.295±0.005	0.275±0.008
a8a	22696	0.702±0.002	**0.664±0.002**	**0.664±0.002**	0.664±0.002	0.664±0.002
w7a	24692	0.291±0.002	**0.283±0.002**	**0.283±0.002**	0.284±0.002	0.284±0.002
a9a	32561	0.698±0.005	**0.664±0.000**	**0.664±0.002**	**0.664±0.000**	**0.664±0.002**
ijcnn1	49990	0.434±0.005	**0.389±0.002**	**0.389±0.002**	0.393±0.001	0.463±0.001
cod-rna	59535	0.686±0.002	/	/	0.614±0.001	0.614±0.001
connect-4	67757	0.781±0.015	/	/	**0.739±0.002**	**0.739±0.002**
skin_nonskin	245057	3.119±0.023	/	/	**2.620±0.043**	**2.620±0.043**
YearPred	463715	0.198±0.001	/	/	**0.187±0.001**	**0.187±0.001**

速手段组合的高效性以及较少的泛化性能损失。

- 结合 Nyström 采样、PCG 加速的半监督核岭回归 (Nyström–PCG)[135]。
- 结合分治算法、随机特征近似核函数的核岭回归 (DC–RF)[188]。

6.2.1　结合 Nyström 采样、PCG 加速的半监督核岭回归

本节使用表 5.4 中的对比方法进行实验，包括最小二乘回归 (RLS)、半监督核岭回归 (LapRLS)、使用 Nyström 采样的 LapRLS (Nyström)，并使用矩阵求逆 (Direct)、共轭梯度下降 (CG)、预处理共轭梯度下降 (PCG) 等不同求解方式。各对比方法的时间复杂度、空间复杂度如表 5.4 所示。

6.2.1.1　测试误差、训练时间对比

首先使用 10 折交叉验证选择最优核函数。其次，重复 30 次将训练数据以 70%、30% 的随机比例划分为训练集、测试集。随机取 10% 数据作为有标签数据 ($n = 0.1(n+u)$)， 10% 数据作为 Nyström 采样点 ($s = 0.1(n+u)$)。

表 6.6 记录了测试集上的均方根误差 (RMSE)，而表 6.7 记录了训练所需的平均迭代次数、平均训练时间（秒）。从表 6.6、表 6.7 可以总结出

(1) RLS 在所有数据上测试误差最高，而 LapRLS-CG、LapRLS-PCG 在所有数据集上测试误差最低。

(2) LapRLS 方法与 Nyström 近似 LapRLS 方法的测试误差无显著差异。

表 6.7 Nyström-PCG 相关算法迭代次数、训练时间对比

	RLS-CG		LapRLS-CG		LapRLS-PCG		Nyström-CG		Nyström-PCG	
	iter	time	iter	time	iter	time	iter	time	iter	time
madelon	32	0.003	13	0.029	6	0.032	12	0.043	**1**	**0.006**
space_ga	11	0.004	23	1.220	5	0.569	23	0.113	**2**	**0.016**
abalone	64	0.053	98	26.50	4	0.903	94	0.363	**2**	**0.067**
phishing	74	0.031	300	24.20	56	8.210	300	2.470	**3**	**0.045**
a8a	100	0.068	50	189.1	3	20.98	50	44.71	**1**	**4.370**
w7a	13	0.072	32	143.2	2	9.683	213	107.7	**1**	**2.252**
a9a	300	0.529	64	1699	3	30.30	65	70.40	**1**	**4.034**
ijcnn1	242	8.204	57	2154	9	72.41	53	108.8	**5**	**4.186**
cod-rna	96	7.178	/	/	/	/	55	134.6	**7**	**8.154**
connect-4	103	11.07	/	/	/	/	154	186.5	**10**	**4.220**
skin_nonskin	43	91.39	/	/	/	/	65	1490	**3**	**40.05**
YearPred	37	236.5	/	/	/	/	94	2479	**2**	**116.1**

(3) 由于内存限制，LapRLS-CG、LapRLS-PCG 无法在大规模数据上运行。

(4) 使用 CG、PCG 求解闭式解得到的测试误差相似，但使用 PCG 方法的迭代次数比使用 CG 方法的迭代次数少很多、计算效率高很多。

(5) Nyström-PCG 达到了与 LapRLS 类似的泛化性能，但在计算效率上带来成百上千倍的加速。

6.2.1.2 有标签数据比例的影响

为探究标签比例对泛化性能的影响，调整有标签数据个数 $n \in (n+u) \times \{1\%, 2\%, 4\%, 8\%, 16\%, 32\%, 64\%\}$，同时固定 Nyström 采样数为 $s = 0.1(n+u)$。使用矩阵运算、CG、PCG求解闭式解，对泛化性能影响不大，因此只关心对应闭式解形式 RLS、 LapRLS、Nyström 近似 LapRLS （缩写为 Nyström）对应的测试误差。通过不同有标签 / 无标签数据上重复 20 次实验，将平均 RMSE 、标准差报告在图 6.3，可以看出

(1) LapRLS 方法的测试误差总是小于 RLS 方法的测试误差，但两者测试误差的差异随着有标签数据 n 的增多而减小。

(2) 当有标签数据足够多时，LapRLS、LapRLS–Nyström 平均精度很接近。

(3) 所有方法测试误差的标准差都随着有标签样本个数 n 的增加而减小。

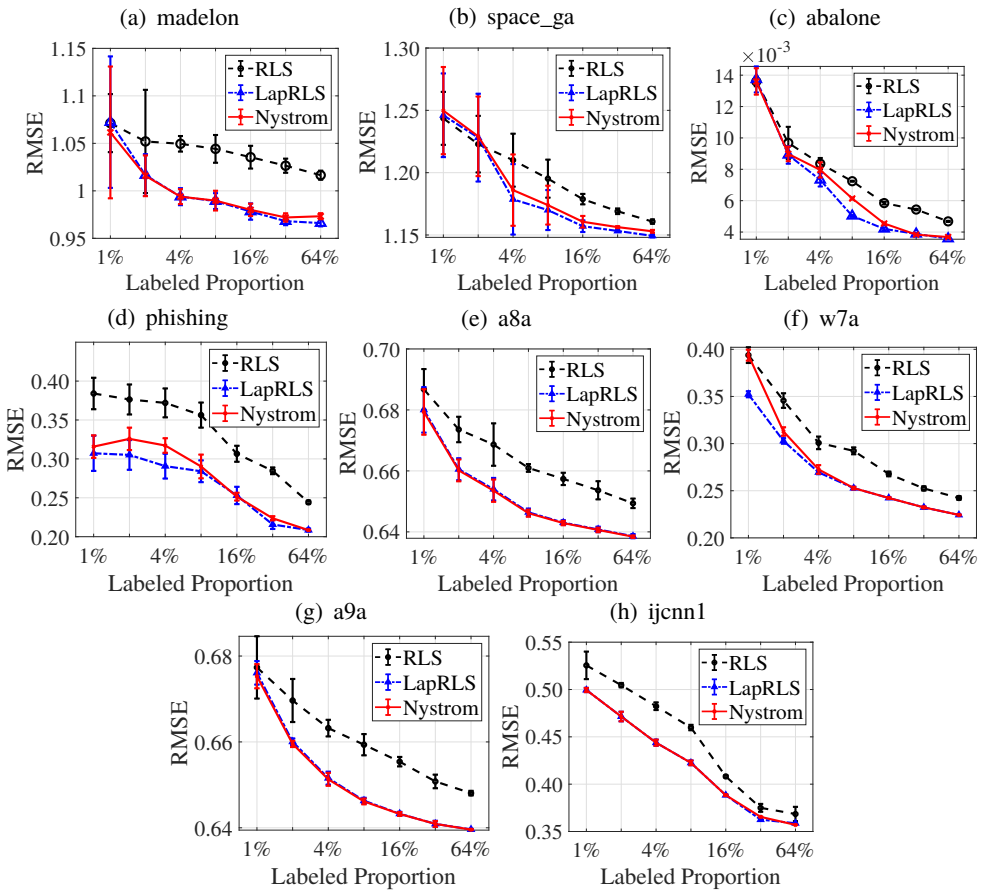

图 6.3 不同标签比例对测试误差 (RMSE) 的影响

6.2.1.3　Nyström 采样比例的影响

为探究 Nyström 采样比例对泛化性能的影响，令采样数据动态变化 $s \in (n+u)\times\{1\%, 2\%, 4\%, 8\%, 16\%, 32\%, 64\%\}$，同时固定有标签数据规模为 $n = 0.1(n+u)$。在对 8 个数据集进行不同比例的标记/未标记划分，并重复的 20 次实验，将每种方法的平均泛化误差汇总在图 6.4 中。图 6.4 说明

(1) 因为在相同的带标签/不带标签的分块上运行，RLS 和 LapRLS 方法总是给出相同的测试误差。

(2) Nyström 近似的 LapRLS 方法的测试精度随着采样比例的上升而提高。

(3) 当采样比例大于 10% 之后，Nyström 近似 LapRLS 方法的泛化性能与 LapRLS 相似。

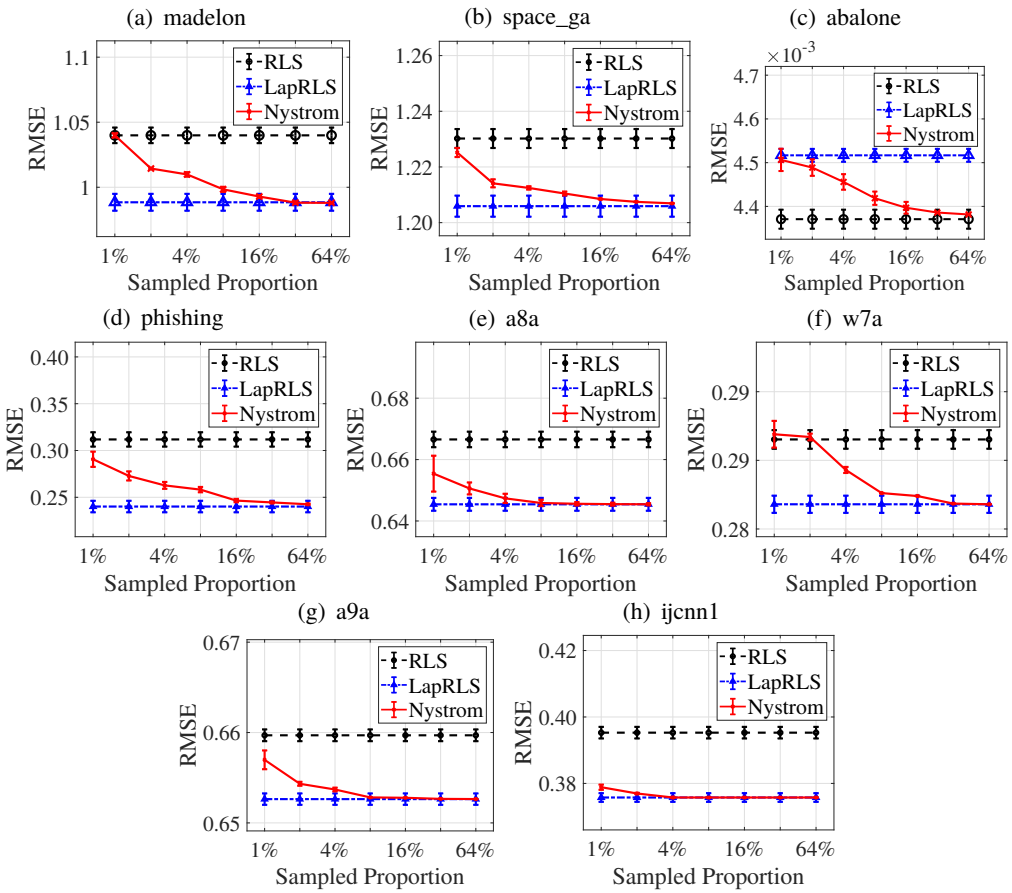

图 6.4 不同 Nyström 采样比例对测试误差 (RMSE) 的影响

6.2.2 结合分治算法、随机特征的核岭回归算法

6.2.2.1 模拟数值实验

从定理 3.16，可知最优泛化误差收敛率为

$$\mathcal{O}\left(n^{-\frac{2r}{2r+\gamma}}\right) = \mathcal{O}\left(n^{-\frac{1}{2r/\gamma+1}}\right)$$

因此，泛化误差收敛率随着比率 $\frac{r}{\gamma}$ 增大而变快。由于 $r \in [1/2, 1]$、$\gamma \in [0, 1]$, 最难的问题对应于 $\frac{r}{\gamma}$ 的最小比率，此时 $r = 1/2, \gamma = 1$，泛化误差收敛率为 $\mathcal{O}(n^{-0.5})$。同时，简单问题对于更大比率 $\frac{r}{\gamma}$，也就是更快的泛化误差收敛率。借鉴研究工作[74,255] 中数值实验中模拟数据的构造方法，使用 $q \geq 2$ 的样条核函数 (spline kernel) 构造模拟数据（研究工作[256] 中等式 2.1.7 对样条函数进行了详细介绍）

$$\Lambda_q(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k\in\mathbb{Z}} \frac{e^{2\pi i\kappa(\boldsymbol{x}-\boldsymbol{x}')}}{|k|^q} = 1 + 2\sum_{k=1}^{\infty} \frac{\cos(2\pi k(\boldsymbol{x}-\boldsymbol{x}'))}{k^q}. \tag{6.1}$$

其中关键在于，样条函数能够天然地构造随机特征。$\forall\, q, q' \in \mathbb{R}$，存在

$$\int_0^1 \Lambda_q(\boldsymbol{x}, z)\Lambda_{q'}(\boldsymbol{x}', z)dz = \Lambda_{q+q'}(\boldsymbol{x}, \boldsymbol{x}'). \tag{6.2}$$

使用以下设置，从而对对简单问题和困难问题进行实验

- **输入分布**: $\mathcal{X} = [0,1]$。对应边际分布 ρ_X 为均匀分布。
- **输出分布**: 目标函数为 $f_*(\boldsymbol{x}) = \Lambda_{\frac{r}{\gamma}+\frac{1}{2}}(\boldsymbol{x}, 0)$，并有方差 ϵ。
- **核函数及随机特征**: 核函数定义为 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \Lambda_{\frac{1}{\gamma}}(\boldsymbol{x}, \boldsymbol{x}')$。根据随机特征定义、等式 (6.2) 可得，$\psi(\boldsymbol{x}, \omega_i) = \Lambda_{\frac{1}{\gamma}}(\boldsymbol{x}, \omega_i)$，其中 ω_i 独立同分布地采样于均匀分布 $U[0,1]$。样条核函数对应随机特征为

$$\phi(\boldsymbol{x}) = M^{-1/2}\big(\psi(\boldsymbol{x}, \omega_1), \cdots, \psi(\boldsymbol{x}, \omega_M)\big).$$

当比率 $\frac{r}{\gamma}$ 接近 0.5，此时问题较难，泛化误差收敛率接近于 $\mathcal{O}(1/\sqrt{n})$。此时，目标函数接近 $\Lambda_1(\boldsymbol{x}, \boldsymbol{x}')$，对应的输出值变化极大、曲线陡峭，因此构造出的数据难以拟合。当比率 $\frac{r}{\gamma}$ 远离 0.5 时，此时问题变容易，泛化误差收敛率逐渐接近于 $\mathcal{O}(1/n)$。此时，输出值变化幅度较小、输出曲线越来越平缓，因此构造的数据容易拟合。本节，为 r、γ 设置不同取值，构造出难易程度不同的回归问题。

使用等式 (5.2) 定义的 KRR-DC-RF 学习器 $\widehat{f}_D^M$，并使用不同的样本数 $n \in \{100, \cdots, 10000\}$。同时，根据定理 3.16，将随机特征数设置为 $M = n^{\frac{(2r-1)\gamma+1}{2r+\gamma}}$、分片数设置为 $m = n^{\frac{2r-1}{2r+\gamma}}$。对正则化系数 λ_A 在 $n^{-\frac{1}{2r+\gamma}}$ 附近进行调参。

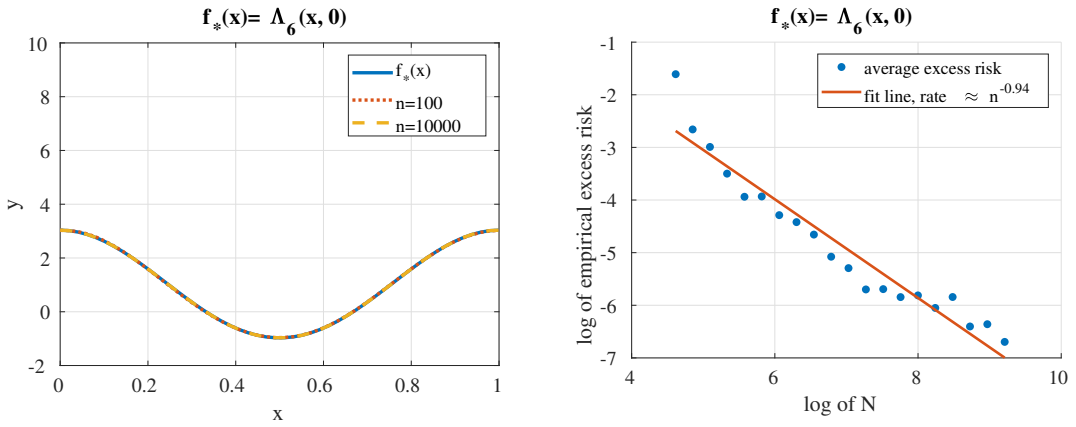

图 6.5 KRR-DC-RF 在简单回归问题上的表现

(1) 简单问题

使用 $r = 11/16$、$\gamma = 1/8$，并将方差设置为 $\epsilon = 0.1$，从而构造出较为简单的回归问题。此时目标函数为 $f_*(\boldsymbol{x}) = \Lambda_6(\boldsymbol{x}, 0)$，样本输出曲线平缓，容易拟合。图 6.5 中左边子图显示，只要少量数据 $n = 100$ 就能拟合很好；而右边子图显示，泛化误差的最佳拟合直线对应斜率为 −0.94，与通过最优泛化理论计算出来的泛

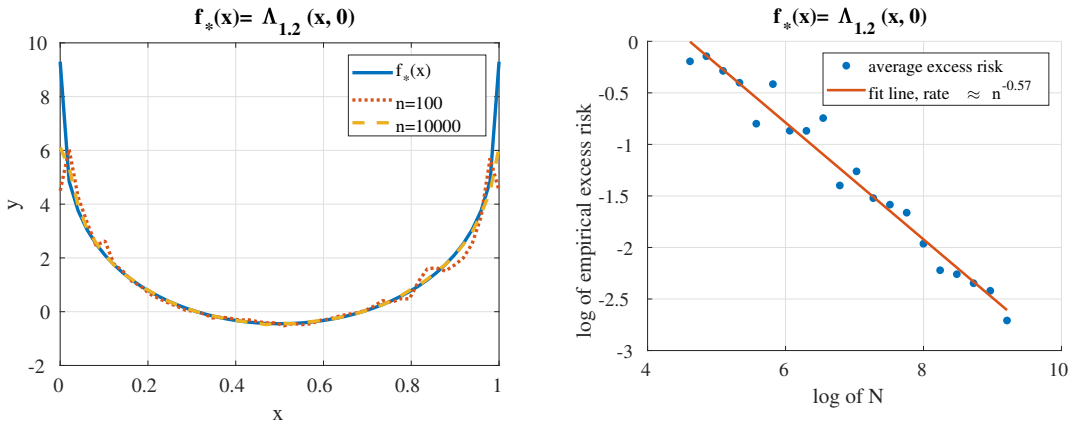

图 6.6 KRR-DC-RF 在困难回归问题上的表现

化误差收敛率 $\mathcal{O}(n^{-0.92})$ (定理 3.16) 非常相近。因此在容易的问题上的测试结果（泛化误差收敛率较快），验证了 KRR-DC-RF 的泛化理论结果。

(2) 困难问题

使用 $r = 0.7$、$\gamma = 1$、$\epsilon = 0.1$，并将方差设置为 $\epsilon = 0.1$，从而构造出较为困难的回归问题。此时目标函数为 $f_*(\boldsymbol{x}) = \Lambda_{1.2}(\boldsymbol{x}, 0)$，样本输出曲线陡峭，难以拟合。图 6.6 的左边子图显示, 当输入样本 $\boldsymbol{x}$ 接近于 0 或 1 时，输出值变化剧烈，因此需要更多样本才能拟合的较好 ($n = 10000$)；右边子图显示，经验上的泛化误差收敛率为 $\mathcal{O}(n^{-0.57})$，而通过最优泛化理论得出的泛化误差收敛率为 $\mathcal{O}(n^{-0.58})$。因此在困难的问题上的测试结果（泛化误差收敛率较慢），验证了 KRR-DC-RF 的泛化理论结果。

6.2.2.2 真实数据实验

由于真实数据集对应回归问题通常较难，不妨假定真实任务对应于定理 3.16 中的最差情况 ($\gamma = 1$ $r = 1/2$)。此时，随机特征维度为 $M \gtrsim \mathcal{O}(\sqrt{n})$，而分块数为常数 $m \lesssim \mathcal{O}(1)$。如果经验结果优于最坏情况，则理论结果可以得到验证，即最优学习率比最坏情况需要更少的随机特征或更多的分块数。

在二分类数据集 covtype [3]、SUSY [4]、HIGGS [5] 上重复随机采样 $n = 2.5 \times 10^5$ 数据点（此时 $\sqrt{n} = 500$），在运行 KRR-DC-RF 算法。使用随机傅里叶特征 $\kappa(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \phi_M(\boldsymbol{x}), \phi_M(\boldsymbol{x}') \rangle$ 近似高斯核 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2)$[31]。此时随

[3] https://archive.ics.uci.edu/ml/datasets/covertype

[4] https://archive.ics.uci.edu/ml/datasets/susy

[5] https://archive.ics.uci.edu/ml/datasets/higgs

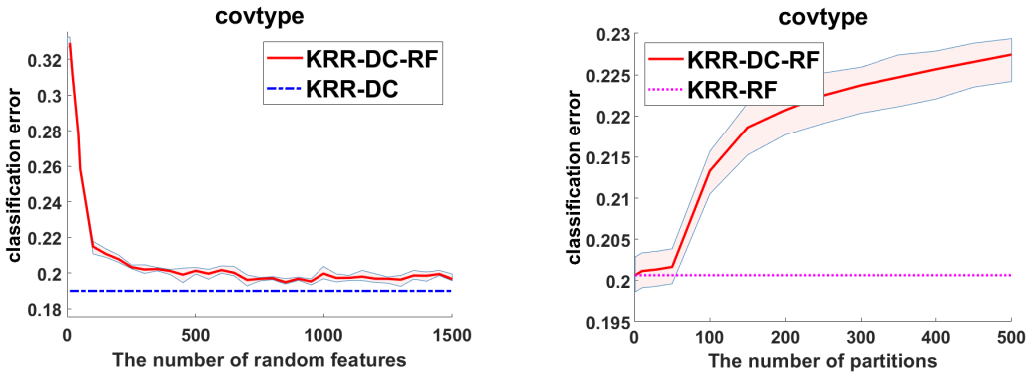

图 6.7 KRR-DC-RF 在 covtype 数据集上的测试误差

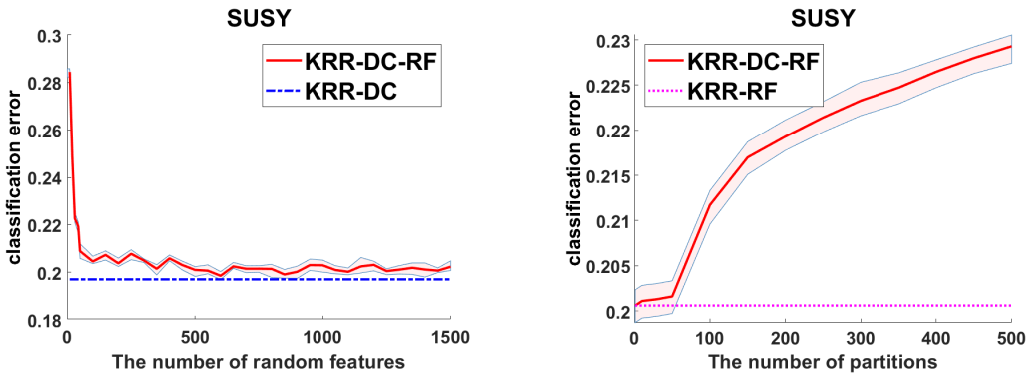

图 6.8 KRR-DC-RF 在 SUSY 数据集上的测试误差

机傅里叶特征为 $\phi_M(\boldsymbol{x}) = \cos(\omega^\top \boldsymbol{x} + b)$，其中 ω 从高斯核对应分布中随机采样，而 b 采样于均匀分布 $[0, 2\pi]$. 在实验中，使用 10 折交叉验证为σ、 λ_A 选取最优参数组合，并重复运行实验 10 次记录平均误差、标准差。

(1) 随机特征维度对泛化性能的影响

将分块数固定为 $m = 20$，并改变随机特征维度 M，从而学习随机特征维度 M 对学习器泛化性能的影响。如图 6.7、图 6.8、图 6.9 的全部左边子图所示，当随机特征维度较小时，KRR-DC-RF 的分类误差随着随机特征维度的增加而减小. 然而，当随机特征维度多于某个阈值后，KRR-DC-RF 收敛，并于 KRR-DC 方法的分类误差相近。图中显示阈值为 $\mathcal{O}(\sqrt{n})$，与理论结果一致（定理 3.16）。

(2) 随机特征维度对泛化性能的影响

将随机特征数固定为 $M = 500$，并改变分块数 m，从而学习分块数 m 对学习器分类误差的影响。如图 6.7、图 6.8、图 6.9 的全部右边子图所示，当分块数为常数级别时，KRR-DC-RF 对应的分类误差较低，与 KRR-RF 非常相近。当分

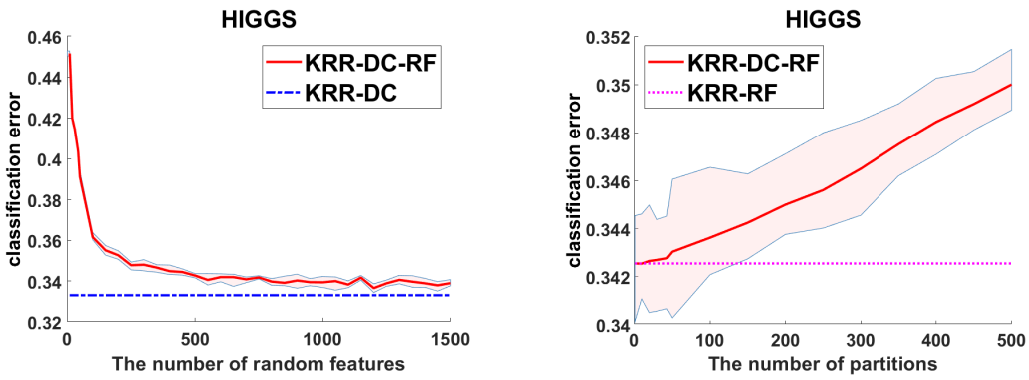

图 6.9 KRR-DC-RF 在 HIGGS 数据集上的测试误差

块数增加时，KRR-DC-RF 对应的分类误差急剧增大，与理论分析中最差情况相吻合 ($\gamma = 1$、$r = 1/2$)。

第7章　总结与展望

核方法模型选择决定了核方法的泛化性能，但现有大规模核方法模型选择方法在大规模半监督数据情形下，在泛化理论、模型选择准则、优化算法三方面均存在不足。本文对半监督核方法泛化理论展开研究，分析影响半监督核方法泛化性能的因素；通过最小化泛化误差上界，指导核方法模型选择准则的制定；结合大规模加速手段，为大规模半监督核方法模型选择的求解设计高效算法。具体研究内容包括：

- **大规模半监督的核方法模型选择理论研究。**基于谱度量、局部 Rademacher 复杂度、积分算子理论，分别给出二分类核方法、半监督核方法、大规模核方法的泛化误差理论。其中，首次为类似于核对齐的方法提供泛化理论保证[206]；通过建立假设空间上 Rademacher 复杂度、损失空间上 Rademacher 复杂度的联系，首次将基于局部 Rademacher 复杂度的泛化理论分析扩展到半监督核方法中[207,208]；通过构造合适的误差分解，首次将积分算子理论的最优泛化理论推广到分布式、随机特征结合的核岭回归方法中[188]。

- **大规模半监督的核方法模型选择准则研究。** 通过最大化谱度量，首次提出了兼具泛化理论保证、高效计算效率的核方法模型选择准则[206]；通过最小化局部 Rademacher 复杂度的泛化误差界，将最小化核矩阵尾部特征值之和与经验损失最小化学习框架 (ERM) 共同训练，并以端到端形式学习核函数[205]；通过界定结合非平稳谱核构造核网络的 Rademacher 复杂度，首次提出将最小化全部数据上特征映射的 Frobenius 范数 $\|\phi(\boldsymbol{X})\|_F^2$ 作为核选择准则，并使用反向传播同时学习核超参数、模型参数[107,108]。

- **大规模半监督的核方法模型选择算法研究。** 为进一步提高核方法求解效率，提出将分布式、随机特征相结合的高效算法，并给出最优泛化理论保证[188]；为提高半监督核岭回归 (LapRLS) 的求解效率，提出将 Nyström 采样、预处理共轭梯度下降算法相结合的方法，并给出泛化理论保证[135]；为求解复杂的多核多分类问题，提出使用对偶梯度下降算法在实数空间更新对偶变量、原变量的范数[205]；为求解最小化权重矩阵尾部奇异值之和等不可微的优化目标，提出使用近端梯度下降算法求解模型参数[207,208]。

未来工作方向为建立核函数与神经网络的关联，主要包括以下几个方面：

(1) 建立核方法与深度神经网络的关联，尝试将核方法中常用的泛化理论分析工具推广到深度神经网络中，为深度神经网络建立泛化误差理论。从而，提高深度神经网络里可解释性，从泛化理论角度指导神经网络设计。

(2) 借鉴深度神经网络结构，通过随机特征显式地构造出核网络。以端到端形式，通过反向传播同时优化核超参数、模型参数。最终，使用神经网络中权重矩阵估计出对应核函数作为模型选择输出的核函数。

(3) 结合神经网络中平均场理论 (mean filed theory)[257,258]，为核函数、核神经网络的初始化超参数提供指导。将谱核网络与神经正切核 (neural tangent kernel, NTK) 结合，达到简化复杂神经网络的目的[259]。

参考文献

[1] Mitchell T M, et al. Machine learning. 1997[J]. Burr Ridge, IL: McGraw Hill, 1997, 45(37): 870-877.

[2] Bishop C M. Pattern recognition and machine learning[M]. springer, 2006.

[3] Vapnik V. The nature of statistical learning theory[M]. Springer Science & Business Media, 1999.

[4] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis[M]. Cambridge University Press, 2004.

[5] Kwok J T, Tsang I. Learning with idealized kernels[C]//Proceedings of the 20th International Conference on Machine Learning (ICML). 2003: 400-407.

[6] Shawe-Taylor J, Cristianini N. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge University Press, 2000.

[7] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3):273-297.

[8] Vapnik V, Chervonenkis A Y. On the uniform convergence of relative frequencies of events to their probabilities[J]. Theory of Probability & Its Applications, 1971, 16(2):264-280.

[9] Rasmussen C E. Gaussian processes in machine learning[C]//Summer School on Machine Learning. Springer, 2003: 63-71.

[10] Saunders C, Gammerman A, Vovk V. Ridge regression learning algorithm in dual variables [J]. 1998.

[11] Schölkopf B, Smola A, Müller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural computation, 1998, 10(5):1299-1319.

[12] Dhillon I S, Guan Y, Kulis B. Kernel k-means: spectral clustering and normalized cuts[C]// Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 551-556.

[13] Stone M. Cross-validatory choice and assessment of statistical predictions[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1974, 36(2):111-133.

[14] Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1-3):131-159.

[15] Efron B, Tibshirani R J. An introduction to the bootstrap[M]. CRC press, 1994.

[16] Wahba G. Support Vector Machine, reproducing kernel Hilbert spaces and the randomized GACV[C]//Advances in Kernel Methods-Support Vector Learning: volume 6. MIT Press, 1999: 69-88.

[17] Vapnik V, Chapelle O. Bounds on error expectation for support vector machines[J]. Neural computation, 2000, 12(9):2013-2036.

[18] An S, Liu W, Venkatesh S. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression[J]. Pattern Recognition, 2007, 40(8):2154-2162.

[19] Debruyne M, Hubert M, Suykens J A. Model selection in kernel based regression using the influence function[J]. Journal of Machine Learning Research, 2008, 9:2377-2400.

[20] Liu Y, Jiang S, Liao S. Efficient approximation of cross-validation for kernel methods using Bouligand influence function[C]//Proceedings of the 31st International Conference on Machine Learning (ICML). 2014: 324-332.

[21] Cristianini N, Shawe-Taylor J, Elisseeff A, et al. On kernel-target alignment[C]//Advances in neural information processing systems. 2002: 367-373.

[22] Cortes C, Mohri M, Rostamizadeh A. Algorithms for learning kernels based on centered alignment[J]. Journal of Machine Learning Research, 2012, 13(Mar):795-828.

[23] Nguyen C H, Ho T B. An efficient kernel matrix evaluation measure[J]. Pattern Recognition, 2008, 41(11):3366-3372.

[24] Baram Y. Learning by kernel polarization[J]. Neural Computation, 2005, 17(6):1264-1275.

[25] Joachims T. Transductive inference for text classification using support vector machines[C]// Proceedings of the 16th International Conference on Machine Learning (ICML): volume 99. 1999: 200-209.

[26] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research, 2006, 7(Nov):2399-2434.

[27] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.

[28] Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates[J]. Journal of Machine Learning Research, 2015, 16 (1):3299-3340.

[29] Lin J, Rosasco L. Optimal learning for multi-pass stochastic gradient methods[C]//Advances in Neural Information Processing Systems 29 (NIPS). 2016: 4556-4564.

[30] Williams C K, Seeger M. Using the nyström method to speed up kernel machines[C]// Advances in Neural Information Processing Systems 14 (NIPS). 2001: 682-688.

[31] Rahimi A, Recht B. Random features for large-scale kernel machines[C]//Advances in Neural Information Processing Systems 21 (NIPS). 2007: 1177-1184.

[32] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines[J]. 1998.

[33] Shalev-Shwartz S, Singer Y, Srebro N, et al. Pegasos: Primal estimated sub-gradient solver for svm[J]. Mathematical programming, 2011, 127(1):3-30.

[34] Cutajar K, Osborne M, Cunningham J, et al. Preconditioning kernel matrices[C]//Proceedings of the 33rd International Conference on Machine Learning (ICML). 2016: 2529-2538.

[35] Aronszajn N. Theory of reproducing kernels[J]. Transactions of the American Mathematical Society, 1950, 68:337-404.

[36] Mercer J. Functions ofpositive and negativetypeand theircommection with the theory ofintegral equations[J]. Philosophicol Trinsdictions ofthe Rogyal Society, 1909:4-415.

[37] Mika S, Ratsch G, Weston J, et al. Fisher discriminant analysis with kernels[C]//Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468). Ieee, 1999: 41-48.

[38] Braun M L. Accurate error bounds for the eigenvalues of the kernel matrix[J]. Journal of Machine Learning Research, 2006, 7(Nov):2303-2328.

[39] Rosasco L, Belkin M, Vito E D. On learning with integral operators[J]. Journal of Machine Learning Research, 2010, 11(Feb):905-934.

[40] Berlinet A, Thomas-Agnan C. Reproducing kernel hilbert spaces in probability and statistics [M]. Springer Science & Business Media, 2011.

[41] Valentini G, Dietterich T G. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods[J]. Journal of Machine Learning Research, 2004, 5(Jul):725-775.

[42] Saunders C, Gammerman A, Vovk V. Ridge regression learning algorithm in dual variables [C]//Proceedings of the 15th International Conference on Machine Learning (ICML). 1998: 515-521.

[43] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3):293-300.

[44] Schölkopf B, Smola A J. Learning with kernels[M]. Cambridge, MA: MIT Press, 2002.

[45] Chang C C, Lin C J. Libsvm: A library for support vector machines[J]. ACM transactions on intelligent systems and technology (TIST), 2011, 2(3):1-27.

[46] Fan R E, Chang K W, Hsieh C J, et al. Liblinear: A library for large linear classification[J]. Journal of machine learning research, 2008, 9(Aug):1871-1874.

[47] Cortes C, Kloft M, Mohri M. Learning kernels using local rademacher complexity[C]// Advances in Neural Information Processing Systems 26 (NIPS). 2013: 2760-2768.

[48] Gai K, Chen G, Zhang C. Learning kernels with radiuses of minimum enclosing balls[C]// Advances in Neural Information Processing Systems 23 (NIPS). 2010: 649-657.

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

[50] Bousquet O, Elisseeff A. Stability and generalization[J]. Journal of Machine Learning Research, 2002, 2(Mar):499-526.

[51] Poggio T, Rifkin R, Mukherjee S, et al. General conditions for predictivity in learning theory [J]. Nature, 2004, 428(6981):419-422.

[52] Zhou D X. The covering number in learning theory[J]. Journal of Complexity, 2002, 18(3): 739-767.

[53] Luxburg U V, Bousquet O, Schölkopf B. A compression approach to support vector model selection[J]. Journal of Machine Learning Research, 2004, 5(Apr):293-323.

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

[55] Shawe-Taylor J, Bartlett P L, Williamson R C, et al. Structural risk minimization over data-dependent hierarchies[J]. IEEE transactions on Information Theory, 1998, 44(5):1926-1940.

[56] Koltchinskii V. Rademacher penalties and structural risk minimization[J]. IEEE Transactions on Information Theory, 2001, 47(5):1902-1914.

[57] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms [M]. Cambridge university press, 2014.

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

[59] Massart P. Some applications of concentration inequalities to statistics[C]//Annales de la Faculté des sciences de Toulouse: Mathématiques: volume 9. 2000: 245-303.

[60] Dudley R M. The sizes of compact subsets of hilbert space and continuity of gaussian processes[J]. Journal of Functional Analysis, 1967, 1(3):290-330.

[61] Srebro N, Sridharan K. Note on refined dudley integral covering number bound[J]. Unpublished results. http://ttic. uchicago. edu/karthik/dudley. pdf, 2010.

[62] Srebro N, Sridharan K, Tewari A. Smoothness, low noise and fast rates[C]//Advances in neural information processing systems. 2010: 2199-2207.

[63] Bartlett P L, Bousquet O, Mendelson S. Localized rademacher complexities[C]//International Conference on Computational Learning Theory. Springer, 2002: 44-58.

[64] Bousquet O, Koltchinskii V, Panchenko D. Some local measures of complexity of convex hulls and generalization bounds[J]. Lecture Notes in Artificial Intelligence, 2002, 2575: 59-73.

[65] Bartlett P L, Bousquet O, Mendelson S, et al. Local rademacher complexities[J]. The Annals of Statistics, 2005, 33(4):1497-1537.

[66] Koltchinskii V, et al. Local rademacher complexities and oracle inequalities in risk minimization[J]. The Annals of Statistics, 2006, 34(6):2593-2656.

[67] Zhang T. Effective dimension and generalization of kernel learning[C]//Advances in Neural Information Processing Systems. 2003: 471-478.

[68] Smale S, Zhou D X. Shannon sampling ii: Connections to learning theory[J]. Applied and Computational Harmonic Analysis, 2005, 19(3):285-302.

[69] Smale S, Zhou D X. Learning theory estimates via integral operators and their approximations [J]. Constructive approximation, 2007, 26(2):153-172.

[70] Caponnetto A, De Vito E. Optimal rates for the regularized least-squares algorithm[J]. Foundations of Computational Mathematics, 2007, 7(3):331-368.

[71] Smale S, Zhou D X. Shannon sampling and function reconstruction from point values[J]. Bulletin of the American Mathematical Society, 2004, 41(3):279-305.

[72] De Vito E, Caponnetto A, Rosasco L. Model selection for regularized least-squares algorithm in learning theory[J]. Foundations of Computational Mathematics, 2005, 5(1):59-85.

[73] Rudi A, Camoriano R, Rosasco L. Less is more: Nyström computational regularization[C]// Advances in Neural Information Processing Systems 28 (NIPS). 2015: 1657-1665.

[74] Rudi A, Rosasco L. Generalization properties of learning with random features[C]//Advances in Neural Information Processing Systems 30 (NIPS). 2017: 3215-3225.

[75] Lin S B, Guo X, Zhou D X. Distributed learning with regularized least squares[J]. The Journal of Machine Learning Research, 2017, 18(1):3202-3232.

[76] Guo Z C, Lin S B, Shi L. Distributed learning with multi-penalty regularization[J]. Applied and Computational Harmonic Analysis, 2017.

[77] Pillaud-Vivien L, Rudi A, Bach F. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes[J]. arXiv preprint arXiv:1805.10074, 2018.

[78] Rudi A, Carratino L, Rosasco L. Falkon: An optimal large scale kernel method[C]//Advances in Neural Information Processing Systems 30 (NIPS). 2017: 3888-3898.

[79] Lin J, Cevher V. Optimal distributed learning with multi-pass stochastic gradient methods [C]//Proceedings of the 35th International Conference on Machine Learning (ICML). 2018: 3098-3107.

[80] Lin J, Cevher V. Optimal rates of sketched-regularized algorithms for least-squares regression over hilbert spaces[J]. arXiv preprint arXiv:1803.04371, 2018.

[81] Carratino L, Rudi A, Rosasco L. Learning with sgd and random features[C]//Advances in Neural Information Processing Systems 31 (NeurIPS). 2018: 10192-10203.

[82] Rudi A, Calandriello D, Carratino L, et al. On fast leverage score sampling and optimal learning[C]//Advances in Neural Information Processing Systems. 2018: 5672-5682.

[83] Kohavi R, Wolpert D H, et al. Bias plus variance decomposition for zero-one loss functions [C]//ICML: volume 96. 1996: 275-83.

[84] Bartlett P L, Boucheron S, Lugosi G. Model selection and error estimation[J]. Machine Learning, 2002, 48:85-113.

[85] Jebara T. Multi-task feature and kernel selection for svms[C]//Proceedings of the twenty-first international conference on Machine learning. 2004: 55.

[86] Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation[J]. Journal of machine learning research, 2004, 5(Sep):1089-1105.

[87] Golub G H, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter[J]. Technometrics, 1979, 21(2):215-223.

[88] Cawley G C, Talbot N L C. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers[J]. Pattern Recognition, 2003, 36(11):2585-2592.

[89] Cristianini N, Elisseeff A, Elisseeff A, et al. On kernel-target alignment[C]//International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001: 367-373.

[90] Bach F, Lanckriet G, Jordan M. Multiple kernel learning, conic duality, and the SMO algorithm[C]//Proceedings of the 21st International Conference on Machine Learning (ICML). 2004: 41-48.

[91] Kloft M, Brefeld U, Sonnenburg S, et al. Lp-norm multiple kernel learning[J]. Journal of Machine Learning Research, 2011, 12:953-997.

[92] Cortes C, Mohri M, Rostamizadeh A. Learning non-linear combinations of kernels[C]// Advances in Neural Information Processing Systems 22 (NIPS). 2009: 396-404.

[93] Gönen M, Alpaydin E. Localized multiple kernel learning[C]//Proceedings of the 25th international conference on Machine learning. 2008: 352-359.

[94] Liu Y, Liao S, Hou Y. Learning kernels with upper bounds of leave-one-out error[C]// Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 2205-2208.

[95] Sonnenburg S, Rätsch G, Schäfer C, et al. Large scale multiple kernel learning[J]. Journal of Machine Learning Research, 2006, 7:1531-1565.

[96] Rakotomamonjy A, Bach F, Canu S, et al. SimpleMKL[J]. Journal of Machine Learning Research, 2008, 9:2491-2521.

[97] Orabona F, Luo J. Ultra-fast optimization algorithm for sparse multi kernel learning[C]// Proceedings of the 28th International Conference on Machine Learning (ICML). 2011: 249-256.

[98] Argyriou A, Hauser R, Micchelli C A, et al. A dc-programming algorithm for kernel selection [C]//Proceedings of the 23rd international conference on Machine learning. 2006: 41-48.

[99] Gehler P V, Nowozin S. Infinite kernel learning[J]. 2008.

[100] Zhang S, Li J, Xie P, et al. Stacked kernel network[J]. arXiv preprint arXiv:1711.09219, 2017.

[101] Samo Y L K, Roberts S. Generalized spectral kernels[J]. arXiv preprint arXiv:1506.02236, 2015.

[102] Remes S, Heinonen M, Kaski S. Non-stationary spectral kernels[C]//Advances in Neural Information Processing Systems 30 (NIPS). 2017: 4642-4651.

[103] Ton J F, Flaxman S, Sejdinovic D, et al. Spatial mapping with gaussian processes and nonstationary fourier features[J]. Spatial statistics, 2018, 28:59-78.

[104] Sun S, Zhang G, Wang C, et al. Differentiable compositional kernel learning for gaussian processes[J]. arXiv preprint arXiv:1806.04326, 2018.

[105] Sun S, Zhang G, Shi J, et al. Functional variational bayesian neural networks[J]. arXiv preprint arXiv:1903.05779, 2019.

[106] Xue H, Wu Z F, Sun W X. Deep spectral kernel learning[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 4019-4025.

[107] Li J, Liu Y, Wang W. Automated spectral kernel learning[C]//Thirty-Four AAAI Conference on Artificial Intelligence. 2020.

[108] Li J, Liu Y, Wang W. Convolutional spectral kernel learning[J]. arXiv preprint arXiv:2002.12744, 2020.

[109] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning[J]. IEEE Transactions on Neural Networks, 2009, 20(3):542-542.

[110] Castelli V, Cover T M. On the exponential value of labeled samples[J]. Pattern Recognition Letters, 1995, 16(1):105-111.

[111] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using em[J]. Machine Learning, 2000, 39(2-3):103-134.

[112] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labelled and unlabelled data[C]//Advances in neural information processing systems. 1997: 571-577.

[113] Chapelle O, Zien A. Semi-supervised classification by low density separation.[C]// Proceedings of the 10th International Conference on Artificial Intelligence and Statistics (AISTATS): volume 2005. Citeseer, 2005: 57-64.

[114] Li Y F, Kwok J T, Zhou Z H. Semi-supervised learning using label mean[C]//Proceedings of the 26th International Conference on Machine Learning (ICML). ACM, 2009: 633-640.

[115] Zhu X, Ghahramani Z, Lafferty J D. Semi-supervised learning using gaussian fields and harmonic functions[C]//Proceedings of the 20th International conference on Machine learning (ICML). 2003: 912-919.

[116] Sindhwani V, Niyogi P, Belkin M. A co-regularization approach to semi-supervised learning with multiple views[C]//Proceedings of ICML workshop on learning with multiple views: volume 2005. Citeseer, 2005: 74-79.

[117] Shahshahani B M, Landgrebe D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon[J]. IEEE Transactions on Geoscience and remote sensing, 1994, 32(5):1087-1095.

[118] Ratsaby J, Venkatesh S S. Learning from a mixture of labeled and unlabeled examples with parametric side information[C]//Proceedings of the eighth annual conference on Computational learning theory. 1995: 412-417.

[119] Cozman F G, Cohen I, Cirelo M. Unlabeled data can degrade classification performance of generative classifiers.[C]//Flairs conference. 2002: 327-331.

[120] Chapelle O, Sindhwani V, Keerthi S S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research, 2008, 9(Feb):203-233.

[121] Li Y F, Zhou Z H. Towards making unlabeled data never hurt[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(1):175-188.

[122] 周志华. 机器学习[M]. 清华大学出版社, 2016.

[123] Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach[J]. Neural computation, 2000, 12(10):2385-2404.

[124] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization[C]//Advances in neural information processing systems. 2005: 529-536.

[125] Li Y F, Zhou Z H. Towards making unlabeled data never hurt[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1):175-188.

[126] Singla A, Patra S, Bruzzone L. A novel classification technique based on progressive transductive svm learning[J]. Pattern Recognition Letters, 2014, 42:101-106.

[127] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts[J]. 2001.

[128] Belkin M, Niyogi P. Semi-supervised learning on riemannian manifolds[J]. Machine Learning, 2004, 56(1-3):209-239.

[129] Chen K, Wang S. Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(1):129-143.

[130] Fergus R, Weiss Y, Torralba A. Semi-supervised learning in gigantic image collections[C]// Advances in Neural Information Processing Systems 22 (NIPS). 2009: 522-530.

[131] Zhang Y M, Huang K, Geng G G, et al. Mtc: a fast and robust graph-based transductive learning method[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26 (9):1979-1991.

[132] Liu W, He J, Chang S F. Large graph construction for scalable semi-supervised learning [C]//Proceedings of the 27th International Conference on Machine Learning (ICML). 2010: 679-686.

[133] Liu W, Wang J, Kumar S, et al. Hashing with graphs[C]//Proceedings of the 28th International Conference on Machine Learning (ICML). Citeseer, 2011: 1-8.

[134] Wang M, Fu W, Hao S, et al. Scalable semi-supervised learning by efficient anchor graph regularization[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1864-1877.

[135] Li J, Liu Y, Yin R, et al. Approximate manifold regularization: scalable algorithm and generalization analysis[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 2887-2893.

[136] Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models.[C]//WACV/MOTION. 2005: 29-36.

[137] Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression[C]//Conference on Learning Theory. 2013: 592-617.

[138] Dai B, Xie B, He N, et al. Scalable kernel methods via doubly stochastic gradients[C]// Advances in Neural Information Processing Systems 27 (NIPS). 2014: 3041-3049.

[139] Gonen A, Orabona F, Shalev-Shwartz S. Solving ridge regression using sketched preconditioned svrg[C]//Proceedings of the 33rd International Conference on Machine Learning (ICML). 2016: 1397-1405.

[140] Li M, Andersen D G, Park J W, et al. Scaling distributed machine learning with the parameter server[C]//Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14): volume 14. 2014: 583-598.

[141] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and Trends® in Machine learning, 2011, 3(1):1-122.

[142] Dean J, Ghemawat S. Mapreduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1):107-113.

[143] Meng X, Bradley J, Yavuz B, et al. Mllib: Machine learning in apache spark[J]. The Journal of Machine Learning Research, 2016, 17(1):1235-1241.

[144] Kraska T, Talwalkar A, Duchi J C, et al. Mlbase: A distributed machine-learning system. [C]//Cidr: volume 1. 2013: 2-1.

[145] Zheng S, Meng Q, Wang T, et al. Asynchronous stochastic gradient descent with delay compensation[J]. arXiv preprint arXiv:1609.08326, 2016.

[146] Meng Q, Chen W, Wang Y, et al. Convergence analysis of distributed stochastic gradient descent with shuffling[J]. arXiv preprint arXiv:1709.10432, 2017.

[147] Zhang Y, Xiao L. Stochastic primal-dual coordinate method for regularized empirical risk minimization[J]. Journal of Machine Learning Research, 2017, 18(1):2939-2980.

[148] Jaggi M, Smith V, Takác M, et al. Communication-efficient distributed dual coordinate ascent [C]//Advances in neural information processing systems. 2014: 3068-3076.

[149] Wangni J, Wang J, Liu J, et al. Gradient sparsification for communication-efficient distributed optimization[C]//Advances in Neural Information Processing Systems. 2018: 1299-1309.

[150] Predd J B, Kulkarni S R, Poor H V. Distributed kernel regression: An algorithm for training collaboratively[C]//2006 IEEE Information Theory Workshop-ITW'06 Punta del Este. IEEE, 2006: 332-336.

[151] Xu C, Zhang Y, Li R, et al. On the feasibility of distributed kernel regression for big data[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11):3041-3052.

[152] Guestrin C, Bodik P, Thibaux R, et al. Distributed regression: an efficient framework for modeling sensor network data[C]//Proceedings of the 3rd international symposium on Information processing in sensor networks. 2004: 1-10.

[153] Do T N, Poulet F. Classifying one billion data with a new distributed svm algorithm.[C]// RIVF. 2006: 59-66.

[154] Alham N K, Li M, Liu Y, et al. A mapreduce-based distributed svm algorithm for automatic image annotation[J]. Computers & Mathematics with Applications, 2011, 62(7):2801-2811.

[155] Hsieh C J, Si S, Dhillon I. A divide-and-conquer solver for kernel support vector machines [C]//Proceedings of the 31st International Conference on Machine Learning (ICML). 2014: 566-574.

[156] Balcan M F, Liang Y, Song L, et al. Communication efficient distributed kernel principal component analysis[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 725-734.

[157] Chen W Y, Song Y, Bai H, et al. Parallel spectral clustering in distributed systems[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(3):568-586.

[158] Wang S. A sharper generalization bound for divide-and-conquer ridge regression[J]. 2019.

[159] Meng Q, Wang Y, Chen W, et al. Generalization error bounds for optimization algorithms via stability.[C]//Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI). 2017: 2336-2342.

[160] Chen T, Giannakis G, Sun T, et al. Lag: Lazily aggregated gradient for communication-efficient distributed learning[C]//Advances in Neural Information Processing Systems. 2018: 5050-5060.

[161] Zhang K, Tsang I W, Kwok J T. Improved nyström low-rank approximation and error analysis [C]//Proceedings of the 25th International Conference on Machine Learning (ICML). ACM, 2008: 1232-1239.

[162] Kumar S, Mohri M, Talwalkar A. Ensemble nystrom method[C]//Advances in Neural Information Processing Systems 22 (NIPS). 2009: 1060-1068.

[163] Li M, Kwok J T Y, Lü B. Making large-scale nyström approximation possible[C]//Proceedings of the 27th International Conference on Machine Learning (ICML). 2010: 631.

[164] Hsieh C J, Si S, Dhillon I S. Fast prediction for large-scale kernel machines[C]//Advances in Neural Information Processing Systems 27 (NIPS). 2014: 3689-3697.

[165] De Brabanter K, De Brabanter J, Suykens J A, et al. Optimized fixed-size kernel models for large data sets[J]. Computational Statistics & Data Analysis, 2010, 54(6):1484-1504.

[166] Gittens A, Mahoney M W. Revisiting the nyström method for improved large-scale machine learning[J]. Journal of Machine Learning Research, 2016, 17(1):3977-4041.

[167] Si S, Hsieh C J, Dhillon I. Computationally efficient nyström approximation using fast transforms[C]//Proceedings of the 33rd International Conference on Machine Learning (ICML). 2016: 2655-2663.

[168] Si S, Hsieh C J, Dhillon I S. Memory efficient kernel approximation[J]. Journal of Machine Learning Research, 2017, 18(1):682-713.

[169] Chen J, Avron H, Sindhwani V. Hierarchically compositional kernels for scalable nonparametric learning[J]. Journal of Machine Learning Research, 2017, 18(1):2214-2255.

[170] Bach F. Sharp analysis of low-rank kernel matrix approximations[C]//Conference on Learning Theory. 2013: 185-209.

[171] Alaoui A, Mahoney M W. Fast randomized kernel ridge regression with statistical guarantees [C]//Advances in Neural Information Processing Systems 28 (NIPS). 2015: 775-783.

[172] Camoriano R, Angles T, Rudi A, et al. Nytro: When subsampling meets early stopping[C]// Artificial Intelligence and Statistics. 2016: 1403-1411.

[173] Tu S, Roelofs R, Venkataraman S, et al. Large scale kernel learning using block coordinate descent[J]. arXiv preprint arXiv:1602.05310, 2016.

[174] Rahimi A, Recht B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning[C]//Advances in Neural Information Processing Systems 22 (NIPS). 2008: 1313-1320.

[175] Le Q, Sarlós T, Smola A. Fastfood-approximating kernel expansions in loglinear time[C]//

Proceedings of the 30th International Conference on Machine Learning (ICML): volume 85. 2013.

[176] Yang J, Sindhwani V, Avron H, et al. Quasi-monte carlo feature maps for shift-invariant kernels[C]//Proceedings of the 31st International Conference on Machine Learning (ICML). 2014: 485-493.

[177] Kar P, Karnick H. Random feature maps for dot product kernels[C]//Artificial Intelligence and Statistics. 2012: 583-591.

[178] Pham N, Pagh R. Fast and scalable polynomial kernels via explicit feature maps[C]// Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 239-247.

[179] Hamid R, Xiao Y, Gittens A, et al. Compact random feature maps[C]//International Conference on Machine Learning. 2014: 19-27.

[180] Li F, Ionescu C, Sminchisescu C. Random fourier approximations for skewed multiplicative histogram kernels[C]//Joint Pattern Recognition Symposium. Springer, 2010: 262-271.

[181] Vedaldi A, Zisserman A. Efficient additive kernels via explicit feature maps[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3):480-492.

[182] Yang J, Sindhwani V, Fan Q, et al. Random laplace feature maps for semigroup kernels on histograms[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 971-978.

[183] Agrawal R, Campbell T, Huggins J H, et al. Data-dependent compression of random features for large-scale kernel approximation[J]. arXiv preprint arXiv:1810.04249, 2018.

[184] Sriperumbudur B, Szabó Z. Optimal rates for random fourier features[C]//Advances in Neural Information Processing Systems 28 (NIPS). 2015: 1144-1152.

[185] Avron H, Clarkson K L, Woodruff D P. Faster kernel ridge regression using sketching and preconditioning[J]. SIAM Journal on Matrix Analysis and Applications, 2017, 38(4):1116-1138.

[186] McWilliams B, Heinze C, Meinshausen N, et al. Loco: Distributing ridge regression with random projections[J]. stat, 2014, 1050:26.

[187] Heinze C, McWilliams B, Meinshausen N. Dual-loco: Distributing statistical estimation using random projections[C]//Artificial Intelligence and Statistics. 2016: 875-883.

[188] Li J, Liu Y, Wang W. Distributed learning with random features[J]. arXiv preprint arXiv:1906.03155, 2019.

[189] Si S, Shin D, Dhillon I S, et al. Multi-scale spectral decomposition of massive graphs[C]// Advances in Neural Information Processing Systems 27 (NIPS). 2014: 2798-2806.

[190] Chang Y W, Hsieh C J, Chang K W, et al. Training and testing low-degree polynomial

data mappings via linear svm[J]. Journal of Machine Learning Research, 2010, 11(Apr): 1471-1490.

[191] Smola A J, Schölkopf B. Sparse greedy matrix approximation for machine learning[J]. 2000.

[192] Morariu V I, Srinivasan B V, Raykar V C, et al. Automatic online tuning for fast gaussian summation[C]//Advances in Neural Information Processing Systems 22 (NIPS). 2009: 1113-1120.

[193] Chen J, Wang L, Anitescu M. A fast summation tree code for matérn kernel[J]. SIAM Journal on Scientific Computing, 2014, 36(1):A289-A309.

[194] Joachims T. Training linear svms in linear time[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 217-226.

[195] Franc V, Sonnenburg S. Optimized cutting plane algorithm for support vector machines[C]// Proceedings of the 25th International Conference on Machine Learning (ICML). ACM, 2008: 320-327.

[196] Chang K W, Hsieh C J, Lin C J. Coordinate descent method for large-scale l2-loss linear support vector machines[J]. Journal of Machine Learning Research, 2008, 9(Jul):1369-1398.

[197] Hsieh C J, Chang K W, Lin C J, et al. A dual coordinate descent method for large-scale linear svm[C]//Proceedings of the 25th International Conference on Machine Learning (ICML). ACM, 2008: 408-415.

[198] Bottou L, Bousquet O. The tradeoffs of large scale learning[C]//Advances in Neural Information Processing Systems 21 (NIPS). 2008: 161-168.

[199] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of machine learning research, 2011, 12(Jul):2121-2159.

[200] Zeiler M D. Adadelta: an adaptive learning rate method[J]. arXiv preprint arXiv:1212.5701, 2012.

[201] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[202] Dennis Jr J E, Schnabel R B. Numerical methods for unconstrained optimization and nonlinear equations: volume 16[M]. Siam, 1996.

[203] Dennis J E, Jr, Moré J J. Quasi-newton methods, motivation and theory[J]. SIAM review, 1977, 19(1):46-89.

[204] Fasshauer G E, McCourt M J. Stable evaluation of gaussian radial basis function interpolants [J]. SIAM Journal on Scientific Computing, 2012, 34(2):A737-A762.

[205] Li J, Liu Y, Yin R, et al. Multi-class learning: From theory to algorithm[C]//Advances in Neural Information Processing Systems 31 (NeurIPS). 2018: 1591-1600.

[206] Li J, Liu Y, Lin H, et al. Efficient kernel selection via spectral analysis[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI). 2017: 2124-2130.

[207] Li J, Liu Y, Yin R, et al. Multi-class learning using unlabeled samples : Theory and algorithm [C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). 2019.

[208] Li J, Liu Y, Wang W. Learning vector-valued functions with local rademacher complexity[J]. arXiv preprint arXiv:1909.04883, 2019.

[209] Steinwart I, Christmann A. Support vector machines[M]. Springer Verlag, 2008.

[210] Gao W, Zhou Z H. On the doubt about margin explanation of boosting[J]. Artificial Intelligence, 2013, 203:1-18.

[211] Koltchinskii V, Panchenko D. Rademacher processes and bounding the risk of function learning[M]. Springer, 2000: 443-457.

[212] Cortes C, Kuznetsov V, Mohri M, et al. Structured prediction theory based on factor graph complexity[C]//Advances in Neural Information Processing Systems 29 (NIPS). 2016: 2514-2522.

[213] Maurer A. A vector-contraction inequality for rademacher complexities[C]//International Conference on Algorithmic Learning Theory. Springer, 2016: 3-17.

[214] Oneto L, Ghio A, Ridella S, et al. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples[J]. Neural Networks, 2015, 65:115-125.

[215] Xu C, Liu T, Tao D, et al. Local rademacher complexity for multi-label learning[J]. IEEE Transactions on Image Processing, 2016, 25(3):1495-1507.

[216] Koltchinskii V, Panchenko D. Empirical margin distributions and bounding the generalization error of combined classifiers[J]. The Annals of Statistics, 2002, 30:1-50.

[217] Lei Y, Dogan U, Binder A, et al. Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms[C]//Advances in Neural Information Processing Systems 28 (NIPS). 2015: 2035-2043.

[218] Maximov Y, Amini M R, Harchaoui Z. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm[J]. Journal of Artificial Intelligence Research, 2018, 61:761-786.

[219] Cortes C, Mohri M, Rostamizadeh A. Multi-class classification with maximum margin multiple kernel[C]//Proceedings of the 30th International Conference on Machine Learning (ICML). 2013: 46-54.

[220] Lei Y, Dogan Ü, Zhou D X, et al. Data-dependent generalization bounds for multi-class classification[J]. IEEE Transactions on Information Theory, 2019, 65(5):2995-3021.

[221] Yu H F, Jain P, Kar P, et al. Large-scale multi-label learning with missing labels[C]//

Proceedings of the 31st International Conference on Machine Learning (ICML). 2014: 593-601.

[222] Smale S, Zhou D X. Estimating the approximation error in learning theory[J]. Analysis and Applications, 2003, 1(01):17-41.

[223] Steinwart I, Hush D, Scovel C. Optimal rates for regularized least squares regression[C]// Proceedings of the 22nd Conference on Learning Theory (COLT 2009). 2009: 79-93.

[224] Lin J, Rosasco L. Optimal rates for multi-pass stochastic gradient methods[J]. The Journal of Machine Learning Research, 2017, 18(1):3375-3421.

[225] Chang X, Lin S B, Zhou D X. Distributed semi-supervised learning with kernel ridge regression[J]. Journal of Machine Learning Research, 2017, 18(1):1493-1514.

[226] Cucker F, Smale S. On the mathematical foundations of learning[J]. Bulletin of the American mathematical society, 2002, 39(1):1-49.

[227] Schölkopf B, Herbrich R, Smola A J. A generalized representer theorem[C]//International conference on computational learning theory. Springer, 2001: 416-426.

[228] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2003, 15(6):1373-1396.

[229] Cristianini N, Shawe-Taylor J, Elisseeff A, et al. On kernel-target alignment[C]//Advances in Neural Information Processing Systems 14 (NIPS). 2001: 367-373.

[230] Cortes C, Mohri M, Rostamizadeh A. Two-stage learning kernel algorithms[C]//27th International Conference on Machine Learning, ICML 2010. 2010: 239-246.

[231] Liu Y, Liao S. Eigenvalues ratio for kernel selection of kernel methods[C]//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI). 2015: 2814-2820.

[232] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[233] Gretton A, Borgwardt K, Rasch M, et al. A kernel method for the two-sample-problem[C]// Advances in neural information processing systems. 2007: 513-520.

[234] Lanckriet G R, Cristianini N, Bartlett P, et al. Learning the kernel matrix with semidefinite programming[J]. Journal of Machine learning research, 2004, 5(Jan):27-72.

[235] Liu X, Li M, Wang L, et al. Multiple kernel k-means with incomplete kernels[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI). 2017: 2259-2265.

[236] Liu X, Zhou S, Wang Y, et al. Optimal neighborhood kernel clustering with multiple kernels [C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI). 2017: 2266-2272.

[237] Lanckriet G R G, Cristianini N, Bartlett P L, et al. Learning the kernel matrix with semidefinite programming[J]. Journal of Machine Learning Research, 2004, 5:27-72.

[238] Shalev-Shwartz S, Tewari A. Stochastic methods for l_1-regularized loss minimization[J]. Journal of Machine Learning Research, 2011, 12:1865-1892.

[239] Orabona F, Luo J. Ultra-fast optimization algorithm for sparse multi kernel learning[C]// Proceedings of the 28th International Conference on Machine Learning (ICML). 2011: 249-256.

[240] Stein M L. Interpolation of spatial data: some theory for kriging[M]. Springer Science & Business Media, 1999.

[241] Gentile C. The robustness of the p-norm algorithms[J]. Machine Learning, 2003, 53(3): 265-299.

[242] Xiao L. Dual averaging methods for regularized stochastic learning and online optimization [J]. Journal of Machine Learning Research, 2010, 11:2543-2596.

[243] Rockafellar R T. Convex analysis[M]. Princeton university press, 1970.

[244] Cai J F, Candès E J, Shen Z. A singular value thresholding algorithm for matrix completion [J]. SIAM Journal on Optimization, 2010, 20(4):1956-1982.

[245] Lu C, Zhu C, Xu C, et al. Generalized singular value thresholding.[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. 2015: 1805-1811.

[246] Parikh N, Boyd S, et al. Proximal algorithms[J]. Foundations and Trends® in Optimization, 2014, 1(3):127-239.

[247] Cawley G C. Leave-one-out cross-validation based model selection criteria for weighted ls-svms[C]//Proceedings of the International Joint Conference on Neural Networks (IJCNN). 2006: 1661-1668.

[248] Nguyen C H, Ho T B. Kernel matrix evaluation[C]//Proceedings of the 20th International Joint Conference on Artifficial Intelligence (IJCAI). 2007: 987-992.

[249] Knerr S, Personnaz L, Dreyfus G. Single-layer learning revisited: a stepwise procedure for building and training a neural network[M]//Neurocomputing. Springer, 1990: 41-50.

[250] Bottou L, Cortes C, Denker J S, et al. Comparison of classifier methods: a case study in handwritten digit recognition[C]//Proceedings of the 12th IAPR International Conference on Pattern Recognition. 1994: 77-82.

[251] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines[J]. Journal of Machine Learning Research, 2002, 2:265-292.

[252] Franc V. Optimization algorithms for kernel methods[J]. Prague: A PhD dissertation. Czech Technical University, 2005.

[253] Zien A, Ong C S. Multiclass multiple kernel learning[C]//Proceedings of the 24th International Conference on Machine Learning (ICML). 2007: 1191-1198.

[254] Huang P S, Avron H, Sainath T N, et al. Kernel methods match deep neural networks on timit

[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing. 2014: 205-209.

[255] Jun K S, Cutkosky A, Orabona F. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration[C]//Advances in Neural Information Processing Systems 33 (NeurIPS). 2019.

[256] Wahba G. Spline models for observational data[M]. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1990.

[257] Poole B, Lahiri S, Raghu M, et al. Exponential expressivity in deep neural networks through transient chaos[C]//Advances in neural information processing systems. 2016: 3360-3368.

[258] Schoenholz S S, Gilmer J, Ganguli S, et al. Deep information propagation[C]//International Conference on Learning Representations. 2017.

[259] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[C]//Advances in neural information processing systems. 2018: 8571-8580.

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

李健，山东省德州人，中国科学院信息工程研究所博士研究生。

已发表(或正式接受)的学术论文:

[1] **Jian Li**, Yong Liu, Weiping Wang. Automated Spectral Kernel Learning. Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). Accepted. (CCF A 类会议).

[2] **Jian Li**, Yong Liu, Rong Yin, Weiping Wang. Multi-Class Learning using Unlabeled Samples: Theory and Algorithm. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2880-2886, 2019. (CCF A 类会议).

[3] **Jian Li**, Yong Liu, Rong Yin, Weiping Wang. Approximate Manifold Regularization: Scalable Algorithm and Generalization Analysis. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2887-2893, 2019. (CCF A 类会议).

[4] **Jian Li**, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, Weiping Wang. Multi-Class Learning: From Theory to Algorithm. Advances in Neural Information Processing Systems 31 (NeurIPS), 1586-1595, 2018. (CCF A 类会议).

[5] **Jian Li**, Yong Liu, Hailun Lin, Yinliang Yue, Weiping Wang. Efficient Kernel Selection via Spectral Analysis. Proceedings of Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2124-2130, 2017. (CCF A 类会议).

预印学术论文:

[1] **Jian Li**, Yong Liu, Weiping Wang. Convolutional Spectral Kernel Learning. arXiv: 2002.12744. (Submission in ICML)

[2] **Jian Li**, Yong Liu, Weiping Wang. Distributed Learning with Random Features. arXiv: 1906.03155. (Submission in JMLR)

[3] Yong Liu*, **Jian Li*** **(equal contribution)**, Lizhong Ding, Weiping Wang. Learning Vector-valued Functions with Local Rademacher Complexity and Unlabeled Data. arXiv: 1909.04883. (Submission in JMLR)

[4] Yong Liu, **Jian Li**, Guangjun Wu, Lizhong Ding, Weiping Wang. Efficient Cross-Validation for Semi-Supervised Learning. arXiv:1902.04768. (Submission in IJCAI)

[5] Yong Liu, **Jian Li**, Weiping Wang. Max-Diversity Distributed Learning: Theory and Algorithms. arXiv:1812.07738, 2018.

参加的研究项目

[1] 自然科学基金项目 (No.61703396): 大规模核方法积分算子谱分析的模型选择方法.

[2] 中科院青促会人才项目 (No.2018YFC0823104).

[3] 信工所引进优秀青年人才项目 (Y7Z0111107).

获奖情况

[1] 2019年度，博士研究生国家奖学金.

[2] 2020年度，“朱李月华”优秀博士研究生奖学金.

[3] 2019年度，中国科学院院长优秀奖.

[4] 2019年度，国科大博士研究生国际合作培养计划.

[5] 2019年度，中国科学院大学“三好学生”荣誉称号.

[6] 2018年度，博士研究生国家奖学金.

[7] 2018年度，中国科学院大学“三好学生”荣誉称号.

[8] 2018年度，信息工程研究所所长优秀奖.

[9] 2018年度，信息工程研究所第二研究室优秀学生.

[10] 2017年度，信息工程研究所第二研究室优秀学生.

致　谢

衷心感谢导师王伟平研究员对我的指导。王老师为我们提供宽松、自由的学习工作环境，让我能够潜心科研。王老师在学术研究方向、日常工作生活中都给予我悉心的指导和无私的帮助，让我在科研道路上快速成长。王老师对我为人处事有着诸多教诲，他对我科研方式方法、表达能力、工作习惯中存在的不足循循善诱并给出非常有帮助的意见，他的言传身教让我受益终身。王老师严谨求实的科学精神、广博的知识储备、开阔的人生视野、敏锐的看问题眼光，永远是我学习的榜样。

衷心感谢指导老师刘勇副研究员在科研工作上的指导、生活上的关心与帮助。刘勇老师帮助我补充数学知识、夯实理论基础，从而引领我进入门槛较高的统计机器学习研究领域。在我的研究工作中，刘老师对我进行了耐心细致的指导，使我逐步走上研究的正轨，我在学术上的每一步成长都饱含刘老师的敦敦教诲和全力支持。刘勇老师参与了我的每一篇学术论文的思路讨论与论文修改，启发着我学术思路的展开与深入，耐心指导我研对每一次论文评审意见的思考与反馈。刘勇老师是既是对我严格要求的良师又是志同道合的益友，非常感谢他对我的宽容与信任，为我创造了宽松的科研环境开展学术工作。刘勇老师对科研工作的热爱，严谨的学术作风，高效的工作习惯，乐观的人生态度以及真挚的人格魅力，让我受益匪浅。

衷心感谢研究所其他老师、同学对我的指导、关怀、帮助，与你们相处非常愉快，让我度过了充实而愉快的研究生阶段。衷心感谢家人、朋友的陪伴与支持。你们的陪伴、支持让我心安，鼓励我走过最困难的阶段。