19004035

上海交通大学学位论文

核函数低维随机特征逼近的广义构建方法及应用研究

姓　　名：**罗钦**
学　　号：119032910073
导　　师：**黄晓霖 副教授**
学　　院：**电子信息与电气工程学院**
学科/专业名称：**控制工程**
申请学位层次：**工程硕士**

2022 年 02 月


19004035

19004035

A Dissertation Submitted to
Shanghai Jiao Tong University for Master/Doctoral Degree

RESEARCH ON GENERALIZED CONSTRUCTION METHOD AND APPLICATION OF LOW DIMENSIONAL RANDOM FEATURE APPROXIMATION OF KERNEL FUNCTION

Author: Qin Luo
Supervisor: Associate Prof. Xiaolin Huang

School of Electronics Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, P.R.China
Feb, 2022


19004035

19004035

摘　要

核学习方法作为利用样本之间的配对信息的重要机器学习方法，具有坚实的理论基础，在分类、回归、降维、聚类等问题中有重要的应用。然而在大规模数据问题中超高的运算存储复杂度及欠佳的灵活性限制了核学习方法在复杂的实际场景中的应用。由于核函数的低维随机特征能够显著地降低核学习方法的时间和空间复杂度，并且可以通过级联结构来增强核学习方法的灵活性，因而其成为近年来核学习领域的研究热点。但是现有的核函数随机特征逼近框架要求核函数满足正定性和平移不变性，而常见的线性核函数、多项式核函数及更广泛的距离度量并不满足这两个性质。此外，在计算存储资源进一步减少情况下，现有的核函数随机特征逼近框架通过减少核函数随机特征的数目来降低存储消耗的方式会严重影响核学习方法的泛化性能。因此，从突破正定性和平移不变限制及计算存储资源有限的角度研究核函数的随机特征逼近方式仍然具有学术和应用价值。

本文工作围绕核函数的低维随机特征这个主题进行展开，涵盖随机特征的适用范围、随机特征的构建方法和随机特征在其他机器学习研究及实际场景中的应用三个方面。本文的主要研究内容如下：

1. 针对现有随机特征逼近框架对核函数的正定性和平移不变性的限制，本文提出了复空间下的不定核的随机特征逼近框架，从理论上证明提出方法的无偏性并且利用正交化的方法减小逼近的方差。最后本文实验验证了相较于现有的不定核与非平移不变核的逼近方法，所提的不定核随机特征逼近方法可以取得更好的核函数的逼近性能。

2. 针对在计算存储资源有限的情况下随机特征数目的减少带来核方法泛化性能下降的问题，本文提出了基于低比特量化的随机特征的深层核学习框架。相比于 32 位浮点数表示，低比特量化的随机特征允许深层核函数在计算存储资源有限的情况下宽度和深度的增加。本文从理论上分析了低比特量化能够带来的模型压缩和运算加速的程度。最后在不同的计算存储资源下实验对比了基于低比特和 32 位比特表示的随机特征的深层核方法，验证了在存储资源有限的情况下低比特轻量化表示的随机特征比 32 位浮点数表示的随机特征具有更好的学习性能。

3. 在应用上，本文从神经正切核（NTK: Neural Tangent Kernel）低

19004035

维逼近的角度思考深度神经网络在小样本上的学习问题。针对深度神经网络在小规模数据集上进行优化的过程中出现的“过拟合”情况，本文基于 NTK 低维假设和神经网络低维训练特性，构造了深度神经网络在小样本学习问题上的低维优化空间，验证了其在小样本学习问题上较优的性能。基于此优化空间，本文提出通过元学习的方法对优化空间进行学习和调整。在具体实验上，基于 NTK 低维假设和元学习方法学习得到的低维优化空间能够有效地缓解深度神经网络在小样本学习问题上出现的“过拟合”情况。

关键词：随机特征，不定核，深层核函数，低比特量化，小样本学习，过拟合

19004035

ABSTRACT

As an important machine learning method utilizing the pair information between samples, kernel method has solid theoretical foundation and a wide application in classification, regression, dimensionality reduction, and clustering. However, the extremely high computational complexity in the large-scale problem and weak flexibility restrict the application of kernel method in more complicated scenarios. Random features are proposed to reduce the computational complexity of the kernel method and enhance the flexibility through the cascade structure. It has become a popular topic recently in the kernel learning research. However, the existing random feature approximation framework requires the kernel function to be positive definite and shift-invariant, while some widely used kernel functions like linear kernel and polynomial kernel do not satisfy these two characteristics. In addition, when computation and storage resources are restricted, directly reducing the dimensionality of random features will degrade the generalization performance of kernel methods. Therefore, the research on random features from the perspective of breaking the restriction of positive definiteness as well as shift-invariance and reducing the computation consumption is valuable both in academic and application scenario.

The thesis is developed around the topic of random features, focusing on three important aspects of random features: the scope of application, the construction, and the application in other machine learning research and practical scenarios for random features. The main research contents include:

1. To break through the restrictions of positive definiteness and shift-invariance for the existing random feature approximation, this thesis proposes a random feature approximation framework for indefinite kernels in complex space. The thesis proves the unbiasedness of the random features for indefinite kernels and proposes the orthogonal method to reduce the approximation variance. Eventually, the experiments verify that the proposed random features can achieve better approximation performance compared

19004035

with the existing approximation methods .

2. Considering the problem that the reduction of the dimension of random features under restricted computation and storage resources leads to the degradation of the generalization performance for kernel method, this thesis proposes a deep kernel learning framework based on low-bit quantized random features. Compared with the 32-bit floating-point representation, the random features with low-bit representation allow the deep kernel to increase its width and depth under restricted computation and storage resources. This thesis theoretically analyzes model compression and calculation acceleration rate brought by low-bit quantization. Eventually, the experiment verifies that the low-bit representation of random features possesses better learning performance than the 32-bit floating-point representation under the restricted computation and storage resources.

3. In terms of application, this thesis considers the few-shot learning problem of deep neural networks from the perspective of the low-dimensional approximation for neural tangent kernel (NTK). To tackle the "overfitting" dilemma of deep neural networks when optimized on small datasets, this thesis constructs the low-dimensional optimization space in few-shot learning problem based on the NTK low-dimensional hypothesis and the low-dimensional training characteristics of the neural network. The optimal space demonstrates its superior performance on few-shot learning problems. Based on this optimal space, this paper proposes to adjust the optimal space through meta-learning. The experiment results indicate that when the deep neural network is trained in the space formulated by NTK low-dimensional hypothesis and meta-learning, the "overfitting" dilemma in the few-shot learning problem could be effectively alleviated.

Key words: random features, indefinite kernel, deep kernel, low-bit quantization, few-shot, overfitting

19004035

目　录

19004035

19004035

第一章　绪论

1.1　研究背景及意义

核学习方法是利用样本之间的配对信息构建的核函数替代高维甚至无穷维的非线性特征映射的一种学习方法。核学习方法易于训练，是过去二三十年之间主流的机器学习方法之一，是人工智能、信息科学、运筹学、统计分析、逼近论等多学科的交叉研究领域。研究学者们在理论、算法和应用等方面进行了长期研究，发展了如图1-1所示的以支持向量机（SVM: support vector machine）[1, 2]、支持向量回归（SVR: support vector regression）[3]、核主成分分析（kPCA: kernel principal component analysis）[4]、核谱聚类（KSC: kernel spectral clustering）[5]、核逻辑回归（KLR: kernel logistic regression）[6]、核岭回归（KRR：kernel ridge regression）[7] 为代表的一系列核学习方法，极大地提升了统计机器学习算法的性能。核学习方法具有坚实的理论保证、良好的实际效果，在诸多场合如数据因果分析[8]、图像分类[9]、目标跟踪[10]、文本匹配[11, 12] 等得到了广泛的应用。

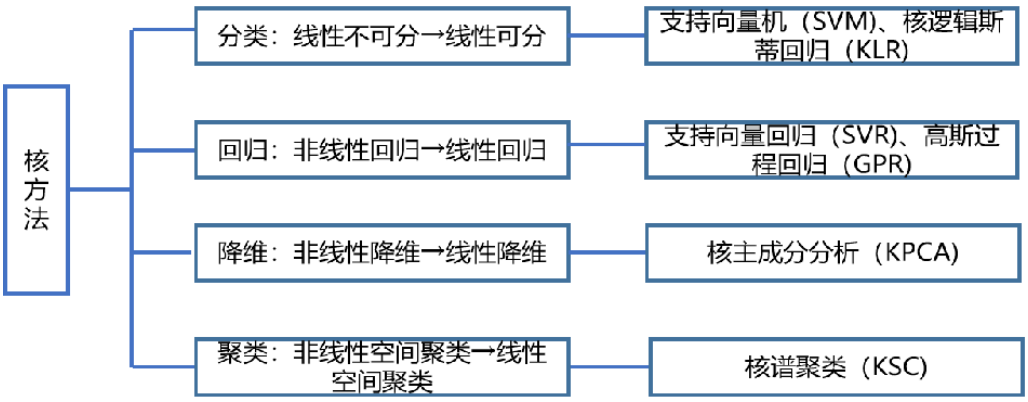


图 1-1　核学习方法在各类学习问题的应用及代表性的机器学习算法

Fig. 1-1　Application of kernel method in various learning problems and its representative machine learning algorithms

纵然传统的核方法在机器学习领域有广泛应用，然而其本身存在两个较大的缺陷：

19004035

1) **大规模核学习问题中时间和空间复杂度大**

传统的核方法时间复杂度和空间复杂度较大，传统核方法严重依赖于核矩阵，而核矩阵的构建以及求逆等运算消耗大量的时间和空间。给定样本数目和样本维度为 N 和 d，SVM、KLR 等核机器学习算法所需要的时间和空间复杂度为 $\mathcal{O}(N^2d)$ 和 $\mathcal{O}(N^2)$。当样本数目 N 极大时，即大规模核学习问题上，传统的核方法消耗时间以及存储成本是难以接受的，严重限制了其在计算资源匮乏的场景中的应用。

2) **传统核学习方法的灵活性欠佳**

传统的核方法需要根据对于应用场景的理解和先验知识人为选定核函数的形式，并通过最大似然估计方法、交叉验证[13] 等方法来确定核函数中的超参数。然而人为选定核函数局限了核方法在实际应用中的性能进一步提升，一方面提供核函数的选择是很有限的，很难遍布所有的核函数空间；另一方面核函数的选择依赖于强烈的先验知识，而难以通过数据驱动的方式，人为选择的核函数往往在给定的具体问题和数据集下并非性能最优的选择。因此传统的核方法在算法的灵活性上欠佳。

大规模核学习[14, 15, 16] 和灵活的核学习[17, 18] 成为了核学习领域研究的两个重点。核函数的低维逼近方法[19, 20, 21] 提升核方法在大规模数据学习问题上的运算效率，降低了存储消耗。通过数据驱动的方式学习核函数的低维逼近中的参数，从而使得在更广泛的函数空间学习到的核函数更加适合给定的数据集和场景任务。随机特征逼近[22, 23] 是核函数的低维逼近最为流行的方法，然而现有随机特征逼近框架大多是针对形式已知的正定核函数设计的，并且不适应于低计算存储资源的环境。因此，研究在更广泛的核函数范畴以及落地场景下的随机特征的构建具有重要的意义。

随机特征及其代表的核函数的低维逼近特性在核方法之外的其他机器学习算法中也有重要的应用，如在目前计算机视觉和语音语义识别领域大热的 Transformer[24, 25] 中，核低维随机特征的采用能够将 Transformer 中的运算量从平方复杂度降低到线性复杂度[26, 27]，从而能够将 Transformer 应用到大尺度的计算机视觉问题与长序列的语义识别问题中。此外随机特征也应用于深度神经网络动态特性[28, 29, 30] 的理解中。尽管深度神经网络在实际应用中的性能超越核方法，然而核方法具有更加坚实的理论基础[31, 32]，从核方法的角度出发能够很好地理解深度神经网络这个“黑盒”的优化过

19004035

程和泛化性能。例如，Belkin 和 Liu 等人[33, 34] 通过调整核函数的随机傅里叶特征的数目来模拟深度神经网络模型复杂度上升的过程，从而发现并解释了深度神经网络中出现的偏差和方差双下降（double descent）的现象。

1.2 国内外研究综述

本节主要对本文研究涉及的基础理论和研究现状进行介绍。1.2.1 节主要介绍核方法中的基本定义及理论，1.2.2 节主要对于核函数的低维随机特征逼近方法的研究现状进行综述。

1.2.1 核方法基础理论

核函数和核矩阵是核方法中最为基础的概念，为了方便后续理解，本小节首先给出核函数和核矩阵的定义。

定义 1.1. （**核函数**）给定任意两个样本 $x,y \in \mathcal{X} \subseteq \mathbb{R}^n$，以及高维甚至无穷维的非线性映射函数 $\phi(.)$，那么两个样本在映射函数作用下的内积可以表示成核函数 $k(.,.) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ 如下：

$$k(x,y) = \langle \phi(x), \phi(y) \rangle \tag{1.1}$$

定义 1.2. （**核矩阵**）给定核函数 $k(.,.) : \mathcal{X}^2 \to \mathbb{R}$，样本 $x_1, x_2, ..., x_N \in \mathcal{X}$，那么对于 $N \times N$ 的矩阵，若矩阵中的每个元素满足：$K_{i,j} := k(x_i, x_j)$，则称该矩阵 K 为关于 k 的核矩阵或 Gram 矩阵。

从定义上看，核函数定义的是两个样本在高维甚至无穷维映射空间之间的相似/不相似的度量，而核矩阵则是数据集中两两样本在高维甚至无穷维映射空间的相似/不相似性的度量矩阵。统计机器学习中有大量利用样本和样本之间的相似信息的算法，如在 k-近邻算法[35] 中通过计算样本与其他样本之间的相似性度量，对于相似度量较大的样本标签进行统计，从而获得该样本的标签。在近年流行的深度学习方法中，相似性度量依然有广泛的应用，如在小样本学习问题中，一类基于度量的方法[36, 37, 38] 也是根据衡量测试集数据和少量训练样本之间的距离来确定样本的标签。在传统的

19004035

核方法中，为了使得使用核函数后优化问题依然保持凸性，约定所采用的核函数为正定核函数。

定义 1.3. （**半正定矩阵**）给定一个实对称矩阵 K，对于任意一个非零向量 α，满足 $\alpha^T K\alpha \geq 0$，那么称 K 是半正定矩阵。K 是半正定矩阵的充分必要条件是矩阵的所有特征值是非负的。

定义 1.4. （**正定核函数**）定义在 $\mathcal{X}\times\mathcal{X}$ 上的核函数 k，如果给定任意的样本数目 N 和所有样本 $x_1,x_2,...,x_N \in \mathcal{X}$，$k$ 产生的 Gram 矩阵为半正定矩阵，即

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(x_i,x_j) \geq 0 \tag{1.2}$$

对任意的 $c \in \mathbb{R}^N$ 成立。则 k 为正定核函数。

如果核函数不满足定义1.4中的正定性条件，则称为不定核。给定一个正定核函数，则其隐式定义的特征空间是一个再生核希尔伯特空间（RKHS: Reproducing Kernel Hilbert Space）。

定义 1.5. （**再生核希尔伯特空间**）定义 $\mathcal{H}$ 是一个由 $f:\mathcal{X}\to\mathbb{R}$ 组成的希尔伯特空间，那么如果存在函数 $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 满足：

1) k 具有再生核性质，即满足对于所有 $f\in\mathcal{H}$，$\langle f,k(x,.)\rangle = f(x)$。

2) k 张成 $\mathcal{H}$。

则称 $\mathcal{H}$ 是具有点积 $\langle .,.\rangle$ 的再生核希尔伯特空间。

从定义1.4判断核函数是否正定需要考虑两个“任意”的条件，一个是任意的样本数目，另一个是任意的样本取值，而这两个”任意“的条件”使得难以从核函数定义的角度去判断核函数是否正定。核方法领域广泛使用的核函数大部分属于 Mercer 核函数这个类别，Mercer 定理和 Mercer 核映射定理给出了 Mercer 核函数的定义与正定性保证。

定理 1.1. （Mercer **定理**[2]）假定 $k\in\mathcal{L}_\infty(\mathcal{X}^2)$ 是满足如下条件的对称实函数：其对应的积分运算 T_k 是正定的，其中 T_k：$\mathcal{L}_2(\mathcal{X})\to\mathcal{L}_2(\mathcal{X})$，$(T_k f)(x) := \int_{\mathcal{X}} k(x,y)f(y)d\mu$，即对于所有 $f\in\mathcal{L}_2(\mathcal{X})$，有

$$\int_{\mathcal{X}^2} k(x,y)f(x)f(y)d\mu(x)d\mu(y) \geq 0 \tag{1.3}$$

19004035

令 $\psi_j \in L_2(\mathcal{X})$ 是 T_k 的单位正交特征函数，相关联的特征值 $\lambda_j \geq 0$ 以非递增的顺序排列，那么有

1) $(\lambda_j)_j \in \ell_1$

2) 对于任意的 $x,y \in \mathcal{X}$，有 $k(x,y) = \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \psi_j(x)\psi_j(y)$ 成立，其中 $N_{\mathcal{H}} \in \mathbb{N}$ 或者 $N_{\mathcal{H}} \in \infty$。

Mercer 定理揭示的是正定 Mercer 核函数可以表示成有限或无穷个特征函数之积的和的形式，这是核函数的低维逼近方法的理论支撑，即可以采用有限维度的特征函数之内积的和来逼近无穷维度的特征函数之内积的和。

定理 1.2. （Mercer **核映射定理**[2]）若 k 是满足定理1.1的核函数，那么对于任意的 $x,y \in \mathcal{X}$，可以构造高维或者无穷维的映射，使得

$$\langle \phi(x), \phi(y) \rangle = k(x,y) \tag{1.4}$$

从定理1.2可以看出，Mercer 核函数具有对应的非线性特征映射函数，结合定义1.4和定理1.2可以推导得到 Mercer 函数的正定性质。

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(x_i,x_j) = \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle = \left\| \sum_{i=1}^{N} c_i \phi(x_i) \right\|^2 \geq 0 \tag{1.5}$$

Mercer 核函数涵盖了部分正定核函数，以下给出常用的正定核函数：

- **高斯核函数**

$$k(x,y) = \exp(-\frac{||x-y||^2}{2\sigma^2}) \tag{1.6}$$

  其中 $\sigma > 0$。σ 为高斯核函数的带宽，σ 越大，核函数的取值随着两个样本之间的欧氏距离的变化越平缓。高斯核函数是核方法中最常用的核函数类型，通过交叉验证方式等方法可以调节其带宽参数使得其在给定任务上性能达到较好水平。

- **拉普拉斯核函数**

$$k(x,y) = \exp(-\frac{||x-y||}{\sigma}) \tag{1.7}$$

19004035

与高斯核函数相同，拉普拉斯核函数的带宽参数 $\sigma > 0$。与高斯核函数不同的是，拉普拉斯核函数对带宽参数 σ 的变化不敏感。

- **径向核函数**（RBF: Radial Basis Function）

$$k(x,y) = \phi(||x-y||_p) \tag{1.8}$$

其中 $p > 0$, ϕ 是实数域上的函数，径向核函数与两个样本之间的 p-范数距离相关。高斯核函数与拉普拉斯核函数是径向核函数的两个特例，对于高斯核函数 $p = 2$，而对于拉普拉斯核函数 $p = 1$。径向核函数具有重要的性质，径向核函数的傅里叶变换是径向函数，后续针对不定核的随机特征构建的研究将会使用到这个性质。

- **线性核函数**

$$k(x,y) = \langle x,y \rangle \tag{1.9}$$

内积和机器学习算法中常用余弦距离的评价指标有紧密联系，当向量 x,y 为单位向量时，线性核函数等价于余弦距离。此时线性核函数的值越大，代表两个向量线性相似程度越大。

- **多项式核函数**

$$k(x,y) = (\langle x,y \rangle + c)^d \tag{1.10}$$

其中 $d \in \mathbb{N}$, $c \geq 0$。当 $d = 1$ 且 $c = 0$ 的时候，多项式核函数可以退化成线性核函数。

- Sigmoid **核函数**

$$k(x,y) = \tanh(\beta \langle x,y \rangle + \theta) \tag{1.11}$$

其中 tanh 为双曲正切函数，$\beta > 0$, $\theta < 0$。Sigmoid 函数是条件正定核，当 $\beta < 0$ 时，Sigmoid 核函数不具有正定性质。

定义 1.6.（**平移不变核函数**）定义在 $\mathcal{X} \times \mathcal{X}$ 上的核函数 k，如果对于任意的样本 x 和 y，核函数 $k(x,y)$ 是两个样本之差的函数，即 $k(x,y) = k(x-y)$，则核函数 $k(x,y)$ 具有平移不变性质。

19004035

根据平移不变核函数的定义，在常用的正定核函数之中，包含高斯核函数、拉普拉斯核函数在内的径向核函数满足平移不变的性质。而线性与多项式核函数是关于样本间内积的函数，不具有平移不变性质，其为非平移不变核。

在常用的正定核函数基础上，可以通过下面的方法生成新的正定核函数：

- 假设 k_1 和 k_2 是正定的核函数，对于所有 $\gamma_1 > 0$ 和 $\gamma_2 > 0$，这两个核函数的线性组合 $\gamma k_1 + \gamma k_2$ 也是正定核函数。
- 假设 k_1 和 k_2 是正定的核函数，这两个核函数的直积 $k_1 \otimes k_2(x,y) = k_1(x,y)k_2(x,y)$ 是正定核函数。
- 假设 k_1 是正定核函数，对于所有函数 $g(x)$，$k(x,y) = g(x)k(x,y)g(y)$ 是正定核函数。

定理 1.3. （**表示定理**[39]）令 $\mathbb{H}$ 为正定核函数 $k(.,.)$ 对应的 RKHS 空间，$||h||_{\mathbb{H}}$ 表示为 $\mathbb{H}$ 空间中关于 h 的范数，给定任意单调递增函数 $\Omega : [0,\infty] \to \mathbb{R}$ 和任意的非负损失函数 $\ell : \mathbb{R}^N \to [0,\infty]$，优化问题

$$\min_{h\in\mathbb{R}} F(h) = \Omega(||h||_{\mathbb{H}}) + \ell(h(x_1), h(x_2), ..., h(x_N)) \tag{1.12}$$

的解总可以写成为：

$$h^*(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i) \tag{1.13}$$

表示定理在核方法中具有重要的作用。采用了核方法后，优化问题(1.12)中的最优解 $h^*(x)$ 可以写成核函数 $k(x,x_i)$ 的线性组合的形式，而对损失函数 $l(x)$ 的形式没有要求，对正则化项的要求仅为单调递增。从表示定理出发，研究者们发展出系列以核函数为中心的机器学习方法，称为核方法或者核学习方法。下面主要介绍本文涉及到的两个核学习方法，支持向量机和支持向量回归算法。

1) **支持向量机算法**

下面介绍的基于带有软间隔的支持向量机算法，图1-2为数据线性可分情况下支持向量机算法示意图，软间隔允许支持向量机算法在线性决策面

19004035

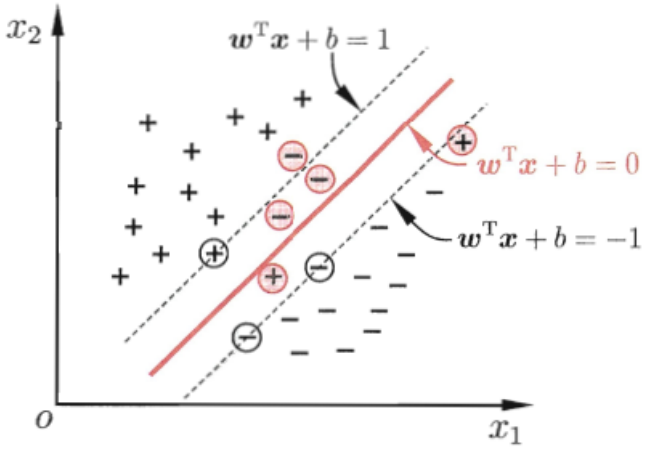


图 1-2　数据线性可分情况下带有软间隔的支持向量机示意图[40]

Fig. 1-2　Support vector machine with soft interval when data is linearly separable [40]

ξ 范围内的样本出错，其中 $\xi \geq 0$。给定训练数据集 $\{x_i, y_i\}_{i=1}^{N}$，与图1-2有所不同的是，大部分情况下训练数据集无法通过一个超平面线性可分，因此需要将原始样本映射到高维或者无穷维的映射空间中，在这个特征空间里面训练样本线性可分。假设高维或者无穷维的映射函数为 $\phi(.)$，则决策函数可以表示为

$$f(x) = w^T \phi(x) + b \tag{1.14}$$

其中 w 和 b 是模型参数，SVM 可以表示成如下的带约束优化问题：

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i \\ s.t. \quad &y_i(w^T\phi(x_i)+b) \geq 1-\xi_i \\ &\xi_i \geq 0 \quad i = 1,2,3...,N \end{aligned} \tag{1.15}$$

优化问题中 ξ_i 为松弛变量，其反映了每个样本违背 $y_i(w^T\phi(x_i)+b) \geq 1$ 这个不等式约束的程度，C 称为正则化参数。如果将约束条件写成惩罚函数

19004035

的形式，则优化问题(1.15)也可以写成：

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \max(0, 1 - y_i(w^T\phi(x_i) + b)) \tag{1.16}$$

公式(1.16)称为 Hinge 损失函数。对于优化问题(1.15)采用 Lagrange 乘子法，则这个问题的 Lagrange 函数可以写成：

$$L(w,b,\alpha,\varepsilon,\mu) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i(1 - \xi_i - y_i(w^T\phi(x_i) + b)) - \sum_{i=1}^{N}\mu_i\xi_i \tag{1.17}$$

其中 $\alpha_i \geq 0$，$\mu_i \geq 0$ 是 Lagrange 乘子。对于 w,b,ξ_i 进行求偏导为零，则有：

$$\begin{cases} \dfrac{\partial L}{\partial w} = w - \sum\limits_{i=1}^{N}\alpha_i y_i \phi(x_i) = 0 \\ \dfrac{\partial L}{\partial b} = \sum\limits_{i=1}^{N}\alpha_i y_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = C - \alpha_i - \xi_i = 0 \end{cases} \tag{1.18}$$

代入(1.15)，则可以得到对偶问题

$$\begin{aligned} &\min_{\alpha} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i, x_j) \\ s.t. \quad &\sum_{i=1}^{N}\alpha_i y_i = 0 \\ &0 \leq \alpha_i \leq C \quad i = 1,2,3...,N \end{aligned} \tag{1.19}$$

其中 $k(x_i,x_j) = \langle\phi(x_i),\phi(x_j)\rangle$ 是正定核函数。对偶问题可以通过 SMO 算法进行求解，求解后得到分类面如下：

$$w = \sum_{i=1}^{N}\alpha_i y_i \phi(x_i) \tag{1.20}$$

19004035

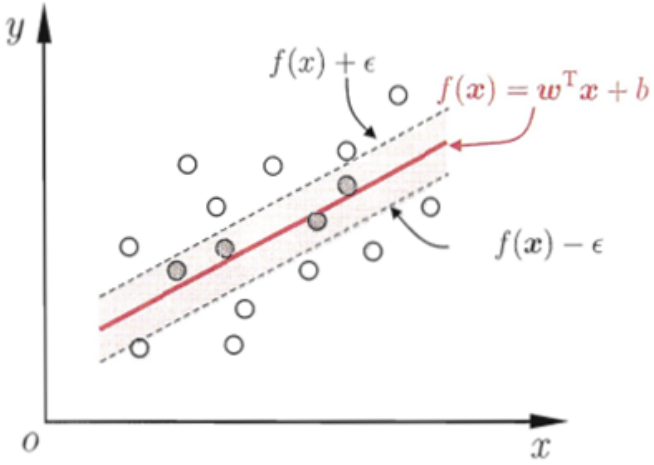


图 1-3　线性相关情况下带有 ε-间隔带的支持向量回归示意图[40]

Fig. 1-3　Support vector regression with ε-interval band in the case of linear correlation [40]

对应的决策函数可以写成：

$$f(x)=\sum_{i=1}^{N}\alpha_i y_i k(x_i,x)+b \tag{1.21}$$

2）支持向量回归算法

下面考虑的是回归问题，给定训练数据集 $D=\{(x_1,y_1),(x_2,y_2)...,(x_N,y_N)\}$，需要学习回归模型 $f(x)$ 使得其与 y 之间的距离尽可能小。如果自变量和因变量之间线性相关，那么采用的回归模型为线性回归，即：

$$f(x)=w^T x+b \tag{1.22}$$

实际大部分情况下自变量和因变量之间并非线性相关，因此和支持向量机方法类似，首先将训练样本通过映射函数 $\phi(.)$ 投射到线性相关的空间中，然后进行线性回归。图1-3是支持向量回归方法，与传统的回归方法不同的是，支持向量回归允许 $f(x)$ 和 y 之间有 ε 的误差，在 2ε 范围内的样本不计算损失误差。则 SVR 的原问题可以写作：

19004035

$$\begin{aligned}
&\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\
&\text{s.t. } f(\phi(x_i)) - y_i \leqslant \varepsilon + \xi_i \\
&\qquad y_i - f(\phi(x_i)) \leqslant \varepsilon + \hat{\xi}_i \\
&\qquad \xi_i \geqslant 0,\ \hat{\xi}_i \geqslant 0, i = 1,2,\ldots,N
\end{aligned} \tag{1.23}$$

ξ_i 和 $\hat{\xi}_i$ 是松弛变量，反映其偏离 ε-间隔带的程度，二者为零时代表样本点在 2ε 区间内。优化问题(1.15)中的约束条件写成惩罚函数的形式，即

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \ell_\varepsilon(f(x_i) - y_i) \tag{1.24}$$

其中 ℓ_ε 定义为：

$$\ell_\varepsilon(z) = \begin{cases} 0, & \text{if } |z| \leqslant \varepsilon \\ |z| - \varepsilon, & \text{otherwise} \end{cases} \tag{1.25}$$

和支持向量机问题类似，对于公式(1.15)所示的 SVR 优化问题，可以通过构造拉格朗日函数并对 w, b, ξ_i, $\hat{\xi}_i$ 求导，从而得到对偶问题：

$$\begin{aligned}
&\max_{\alpha,\hat{\alpha}} \sum_{i=1}^{N} y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon(\hat{\alpha}_i + \alpha_i) \\
&\qquad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\hat{\alpha}_i - \alpha_i)\left(\hat{\alpha}_j - \alpha_j\right)k(x_i, x_j) \\
&\text{s.t. } \sum_{i=1}^{N}(\hat{\alpha}_i - \alpha_i) = 0 \\
&\qquad 0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C
\end{aligned} \tag{1.26}$$

同样地，α_i 和 $\hat{\alpha}_j$ 是对偶变量，$k(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$ 是正定核函数。对偶问题求解后得到对偶变量 α_i 和 $\hat{\alpha}_j$ 的值，则最优解为：

$$w = \sum_{i=1}^{N}(\hat{\alpha}_i - \alpha_i)\phi(x_i) \tag{1.27}$$

19004035

最终回归模型为：

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\,k(x,x_i)+b \tag{1.28}$$

1.2.2 核函数低维随机特征逼近方法综述

核函数低维随机特征的提出的主要目的是解决大规模核学习和灵活的核学习问题，研究者为解决这两个核学习关键问题进行了大量探索。针对大规模的核学习问题，一种简单的思路是采用分治法将数据聚类成 t 个类别，然后对每个类别的数据求解子问题，从而能够逼近原始问题的解[41, 42]。分治法能够将核方法的时间复杂度降低为 $\mathcal{O}(N^2/t)$，空间复杂度降低为 $\mathcal{O}(N^2/t^2)$。由于该方法的时间复杂度依然为 $\mathcal{O}(N^2)$ 的量级，算法加速效果有限。核低维逼近算法是解决大规模的核学习问题的另一种途径。现有的核低维逼近算法分为两大类，一类是以 Nystrom 采样为代表的对于核矩阵进行低秩逼近的算法，另一种以随机特征为代表的对核函数进行逼近的方法。Nystrom 采样方法[43, 44, 45] 通过随机选取数据样本中的子集，构成原有核矩阵的子矩阵 $K_{m,m}$，来替代原有核矩阵 $K_{N,N}$ 中的信息，其中原有的核矩阵可以近似表示成 $K_{N,N}\approx K_{N,m}K_{m,m}^{-1}K_{m,N}$。Nystrom 方法的时间复杂度为 $\mathcal{O}(Nm^2)$，空间复杂度为 $\mathcal{O}(Nm)$，相较于原始的核学习方法，时间和空间复杂度有了显著下降。区别于对于核矩阵进行逼近的算法，随机特征逼近主要利用了针对特定的核函数的级数展开，选取低阶级数从而通过有限维来逼近无穷维。根据选取的基函数不同，研究者们提出了多种类型的随机特征，如：采用了泰勒展开的随机麦克劳林特征（RM: Random Maclaurin）[46] 和张量素描特征（TS: Tensor Sketch）[47, 48]、采用了傅里叶基函数的随机傅里叶特征（RFF：Random Fourier Features）[22] 以及采用勒让德多项式的随机特征[49] 等。其中随机傅里叶特征是核学习方法中最常使用的随机逼近方法，其使得能够在原问题中利用梯度下降等方法直接求解 SVM、KRR 等。假设随机特征的数目为 s（$s\ll N$），核学习算法的计算复杂度降低为 $\mathcal{O}(Ns^2)$，空间复杂度降低为 $\mathcal{O}(Ns)$。时间复杂度和空间复杂度与样本数目成线性关系，使得核方法能够应用于大规模数据的问题。

针对更加灵活的核学习方法，研究者们主要开展了广度核学习和深层核学习两个方向的工作。广度核学习的代表是多核学习（MKL: Multiple

19004035

Kernel Learning）[50, 51]，其使用多个常用核函数的线性组合构成最终的核函数，即 $k=\sum_l \alpha_l k_l$。其中 k_l 是给定的核函数，α_t 是组合系数。多核学习可以整合形式不同的核函数以及相同形式但不同参数的核函数，达到扩大学习的核函数的空间的目标。但广度核方法存在的最大问题是对函数复杂度扩充效率比较低，并且在训练过程中不容易协同。深层核方法[52, 53, 54] 利用了多层核函数复合结构，比如定义两层核函数，

$$k(\psi(x_i),\psi(x_j))=\langle\phi(\psi(x_i)),\phi(\psi(x_j))\rangle \tag{1.29}$$

关于深层核函数的实现方式，一种方式[55, 56] 是深层核函数对应的核矩阵 $\widetilde{K}$ 可以写成一个自适应的低秩矩阵 F 和正定核矩阵 K 之间的 Hadamard 积，另一种方式[57, 58] 是利用了核函数随机傅里叶特征和神经网络结构来实现多层随机傅里叶特征的复合。因此，随机傅里叶特征对于增强核函数的灵活度和表示能力方面也扮演着举足轻重的作用。凭借着随机傅里叶特征在大规模及灵活的核学习的重要作用，随机傅里叶特征在 *NIPS*2017 学术会议上获得了时间检验奖 (Test-of-time Award)。

核函数的随机傅里叶特征大体可以分为独立于数据进行采样的随机傅里叶特征以及依赖于数据进行采样的随机傅里叶特征。下面分别介绍关于这两种类型的随机傅里叶特征的研究工作。

1）独立于数据进行采样的随机傅里叶特征的相关工作

独立于数据进行采样的随机傅里叶特征的采样分布是核函数的傅里叶变换，对于不同的机器学习任务采样的分布不会发生变化。逼近误差是独立于数据进行采样的随机傅里叶特征最重要的评价指标，对于这类随机傅里叶特征，一个通用的假设是逼近误差越小，在 SVM、KRR 等核机器学习算法上性能表现越好。由于随机傅里叶特征理论证明上是无偏的，而逼近误差主要是由偏差和方差两部分构成。因此针对这类随机傅里叶特征的研究主线是如何减小逼近方差并且能够利用更少的随机特征来实现时间和空间复杂度的优化。基于独立于数据分布的随机傅里叶特征采样的研究工作围绕着基于蒙特卡洛方法（Monte Carlo）、基于伪蒙特卡洛方法（QMC：Quasi-Monte Carlo）以及基于数值积分方法三个方面展开。假设通过 Monte Carlo 采样得到的权值为 $\omega_1,\omega_2...\omega_s$，得到的采样值矩阵 $W=[\omega_1,\omega_2,...,\omega_s]^T$，初

19004035

始的随机傅里叶特征方法对应的采样值矩阵可以写成：

$$W_{\mathrm{RFF}} = \frac{1}{\sigma} G \tag{1.30}$$

其中 $G \in \mathbb{R}^{s \times d}$ 是一个稠密的高斯随机矩阵，矩阵中的每个分量是从标准正态分布进行采样得到的，σ 是高斯核函数中的方差。已有基于蒙特卡洛采样的随机傅里叶特征构建方法的主要区别在于采样值矩阵构建方式的不同。

- FastFood **方法**[59]　这个方法主要是利用 Hadanard 矩阵来加速高斯随机矩阵的构建，公式(1.30)可以重新写做：

$$W_{\mathrm{Fastfood}} = \frac{1}{\sigma} B_1 H G \Gamma H B_2 \tag{1.31}$$

  其中 H 是 Walsh-Hadamard 矩阵，其计算复杂度为 $\mathcal{O}(d \log d)$。$\Gamma \in \{0,1\}^{d \times d}$，称为置换矩阵。$B_1$、$B_2$ 和 G 分别是具有以下性质的对角矩阵：B_2 矩阵的主对角线元素来自 $\{\pm 1\}$，G 对角线元素来自于标准正态分布，B_1 对角线上每个元素根据 $(B_1)_{ii} = ||w_i||_2 / ||G||_F$ 进行计算。FastFood 方法产生的 RFF 是无偏的，但相对于最初的 RFF 方差会更大。

- p-**模型**[60]　该方法是 FastFood 方法的延伸，其对应的采样值矩阵可以写成：

$$W_{\mathcal{P}} = [g^T P_1, g^T P_2, ..., g^T P_s]^T \tag{1.32}$$

  其中 g 是一个长度为 t 的高斯随机向量，$P_i \in \mathbb{R}^{t \times d}$，其中对角线来源于公式(1.31)，即 $H_{i,1}$、$H_{i,2}$、...、$H_{i,d}$。相较于 FastFood 方法，p-模型的方差以 $\mathcal{O}(1/d)$ 的收敛率接近于传统的 RFF 的方差。

- SCRF(Signed Circulant Random Features) **方法**[61]　这个方法主要是采用循环矩阵来加速随机傅里叶特征的构造过程，其采样值矩阵可以写作：

$$W_{\mathrm{SCRF}} = [\nu \otimes C(\omega_1), \nu \otimes C(\omega_2), ..., \nu \otimes C(\omega_t)]^T \tag{1.33}$$

19004035

其中 Rademacher 向量 $v=[v_1,v_2,...,v_t]^T$ 满足 $\mathbb{P}(v_i=1)=\mathbb{P}(v_i=-1)=1/2$。循环矩阵中的 $C(\omega_i)$ 根据 $\omega_i\sim N(0,\sigma^{-2}I_d)$ 产生。SCRF 方法也是无偏估计并且能够具有和初始 RFF 方法一样的逼近方差。

- NRFF(Normalized Random Fourier Features) **方法**[62] 这个方法先将数据归一化到单位球面上，再求取 RFF。这个方法虽然简单，但是可以实现逼近方差的减小。以高斯核函数为例，设 x,y 为任意两个样本，$r=||x-y||_2$，相比于初始的 RFF，该方法能够减小的逼近方差的值为：

$$\mathbb{V}[\mathrm{NRFF}]-\mathbb{V}[\mathrm{RFF}]=-\frac{1}{4s}e^{-r^2}(3-e^{-2r^2}) \tag{1.34}$$

- ORF(Orthogonal Random Features) **方法**[63] 这个方法主要是将随机特征的采样值矩阵做正交化，从而减小逼近方差。以高斯核函数为例，采用此方法对应的采样值矩阵为：

$$W_{\mathrm{ORF}}=\frac{1}{\sigma}SQ \tag{1.35}$$

其中 Q 是一个正交矩阵，可以通过对公式(1.30)求取 QR 分解获得。S 是一个对角矩阵，每一个对角元素主要服从卡方分布。当数据维度 d 很大时，方差减小的比例可以描述成以下式子：

$$\frac{\mathbb{V}(\mathrm{ORF})}{\mathbb{V}(\mathrm{RFF})}\approx 1-\frac{(d-1)e^{-r^2}r^4}{d(1-e^{-r^2})^2} \tag{1.36}$$

和 FastFood 方法类似，为了提升随机特征的计算效率，ORF 方法中的随机正交矩阵可以由一系列的结构矩阵替代[64]。后续研究工作对 ORF 方法进行了一系列的拓展，Choromanski 等人[65, 66] 将 ORF 方法延伸到了任意正定的径向基函数，并且给出了方差减小的两个充分条件：1) 两个样本 2-范数距离足够小且 $\mathbb{E}[||\omega||_2^4]\leq\infty$；2) d 的值比较大并且 $r\leq\frac{1}{4\sqrt{c}}$。

在经典 Monte-Carlo 采样中，各个权值的采样是相互独立的，因此基于经典 Monte-Carlo 方法采样值不能够均匀地覆盖整个采样空间，采样的效率较低。针对这个问题，QMC 方法[67] 利用了低差异化序列来保证采样值在

19004035

采样空间中分布的均匀性。典型的 QMC 方法最初从被积函数中进行均匀采样，后面通过相应的变换函数进行更进一步重采样。假设 $t_1,t_2...t_s \in [0,1]^d$ 为初始采样值，$\phi^{-1}(.)$ 为变换函数，变换函数通过核函数的对应的傅里叶变换获取，那么对应的 QMC 方法的采样值矩阵可以写做：

$$W_{\mathrm{QMC}} = [\phi^{-1}(t_1), \phi^{-1}(t_2)...\phi^{-1}(t_s)] \tag{1.37}$$

由 QMC 方法延伸出了 SSF(Spherical Structured Feature Maps)[68] 和 MM (Moment Matching)[69] 两种方法，SSF 方法利用在球面上均匀分布的点来逼近平移和旋转不变的核函数，从而降低 QMC 的时间与空间复杂度；而 MM 方法主要在序列生成方法上做了改进，将低差异化序列生成算法改进为矩匹配算法。这三个算法都能够有效地降低初始 RFF 方法的逼近方差，从而能够在相同的误差下采用更少的随机特征，实现更有效的低维逼近。

2）**依赖于数据进行采样的随机傅里叶特征的相关工作**

依赖于数据进行采样的随机傅里叶特征与最后模型在具体任务上的性能有直接的关联性。现有针对这类随机傅里叶特征的构造的研究主要聚焦于三种方法，分别是基于杠杆得分函数（leverage score）的方法、基于核函数匹配对齐（kernel alignment）的方法以及端到端核函数训练（end-to-end）的方法。

- leverage score **方法**[70, 71]　leverage score 方法建立在重要性采样的基础上，重要性方法可以根据自定义的采样分布对数据进行重采样。自定义的采样分布使得随机傅里叶特征的采样不再局限在核函数的傅里叶变换中，而是可以结合具体的机器学习问题进行设计。leverage score 是一种确定重要性采样中采样分布的方式，比如 KRR 问题的 leverage score 设计为：

  $$l_\lambda(\omega_i) = p(\omega_i) z_{p,\omega_i}^T(X)(K + N\lambda I)^{-1} z_{p,\omega_i}(X) \tag{1.38}$$

  其中 $p(.)$ 是核函数的傅里叶变换，$z_{p,\omega_i}^T(X)$ 是所有数据在随机采样权值 ω_i 处的随机傅里叶特征值的集合，K 是数据集合 X 下的核矩阵。$l_\lambda(\omega_i)$ 越大，代表在最终重采样中 ω_i 所占据比例越大。Liu 等人[72] 从 kernel alignment 的角度对 leverage score 进行改进，避免了核矩阵

19004035

求逆，从而提高了核随机特征计算效率。

- kernel alignment **方法**[73, 74] 这个方法在采样得到的大量随机傅里叶特征值的基础上，利用优化的方法选出这些随机傅里叶特征值的子集。其中的优化问题可以写做：

$$\max_{\alpha} \sum_{i,j=1}^{N} y_i y_j \sum_{t=1}^{J} a_t z_p(x_i, \omega_t) z_p(x_j, \omega_t) \tag{1.39}$$

其中 J 是候选的随机傅里叶特征的数目，它的数目大于最终选取的随机傅里叶特征数目 s。α 是权重向量，通过优化权重向量来反映随机采样值 ω_i 对于最终结果的重要性。(1.39)所反映的优化问题的物理意义也可以理解为学习权重向量使得核矩阵能够逼近目标理想核矩阵 yy^T。EE-RFF(Energy-based Exploration of Random Features) 与 CLR-RFF(Random Feature Compression via Coresets) 是在 kernel alignment 框架下的两种变形。

- end-to-end **训练方法** 相较于前面从大量的随机傅里叶特征进行选择的方式，end-to-end 训练方法就是直接对采样权值 $\{\omega_i\}_{i=1,2,..,s}$ 或者采样分布 $p(\omega)$ 进行学习。随机傅里叶特征的训练学习方法主要分为两种类型，一种是两阶段的学习训练策略，即首先学习获取随机特征，然后将它们嵌入核方法进行模型训练；另一种是一阶段的学习训练策略，即同时学习核函数的谱分布以及分类回归模型。对于两阶段的学习训练策略，Li 等人[75] 提出了首先利用隐式生成模型学习核函数的谱分布，然后再利用这些随机特征训练线性模型。对于一阶段的学习训练策略，Yu 等人[76] 主要利用 Hinge 损失函数同时优化采样值矩阵 W 与线性分类器，Wilson 和 Adams 等人[77] 在谱域通过高斯混合模型构建核函数，根据分类回归损失函数直接学习高斯混合模型的超参数来来优化构建的核函数。Xie 等人[57] 和 Fang 等人[58] 将一阶段学习训练策略与深层核函数的构建结合，通过随机傅里叶特征和神经网络来构建深层核函数，优化神经网络的过程即等价于优化多层随机傅里叶特征和线性分类器参数。

19004035

1.2.3 现有随机特征构建框架的主要问题

纵然核函数低维随机特征逼近经过了近二十年左右的发展，在理论和应用研究上取得了很大进展。然而核函数低维随机特征逼近领域依然面临着如下的问题，这些问题也是本文研究的重点。

1) **随机特征逼近拓展到不定核与非平移不变核**

上述提及的核随机傅里叶特征的构建对核函数的形式有两个要求，一个是核函数具有平移不变性；二个是核函数是正定的。然而很多核函数不满足这两个性质，如常见的线性核、多项式核[78] 等不满足平移不变性质。同时核函数也可以视作样本与样本之间的距离度量，类似 SNE[79]、t-SNE[80] 以及 KL 散度[81] 等距离度量也可以用作核函数。这些核函数不一定满足正定且平移不变的条件，却在特定的问题具有比正定且平移不变核函数更好

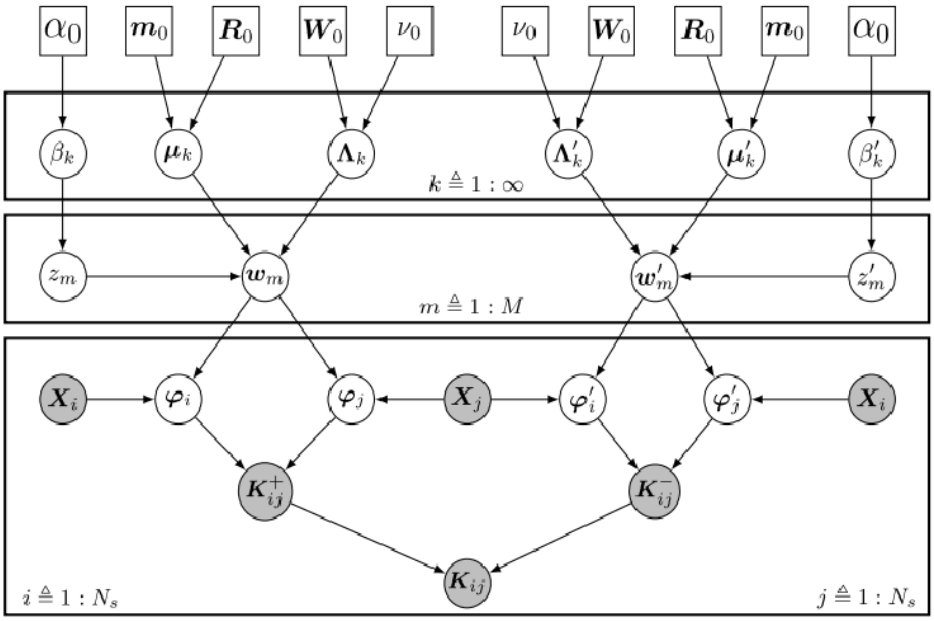


图 1-4 Liu 等人提出的不定核函数随机特征逼近方法示意图[85]

Fig. 1-4 Random features for indefinite kernel proposed by Liu et.al [85]

的性能。因此，研究不定核及非平移不变核的随机特征的构建是必要的。

因为不定核函数及非平移不变核函数的广泛应用和重要意义，近年来

19004035

核函数领域对不定核函数及非平移不变核函数的随机傅里叶特征的构建方法的关注度逐渐升高。关于非平移不变核函数的逼近，本小节前文提到的 RM 和 TS 方法主要针对线性核以及多项式核进行随机逼近。针对加性核函数以及在球面上的高斯核函数的逼近策略，Vedaldi 等人[82] 和 Kafai 等人[83] 给出了相应的解决方案。这些针对非平移不变核的逼近方法虽然是无偏的，但相较于随机傅里叶特征逼近方差较大。关于非正定核函数的逼近，Pennington 等人[84] 发现单位球面上的多项式核函数是不定核，采用正的高斯混合模型模型在频域上逼近不定核的频谱，即：

$$\hat{k}(\omega) = \max\left(0, \sum_{i=1}^{s} c_i \left(\frac{1}{\sqrt{2}\sigma_i}\right)^d e^{-\omega^2/4\sigma_i^2}\right) \tag{1.40}$$

这个方法本质上是用正定核函数逼近不定核函数，这会带来一定程度的逼近误差。Liu 等人[85] 考虑将核矩阵分解成两个正定的核矩阵，并且通过无穷高斯混合模型来获得两个正定核矩阵对应的特征映射，图1-4是其随机特征构建示意图。Pennington 等人和 Liu 等人都采用了高斯混合模型对于非正定核函数进行逼近，然而高斯混合模型对于不定核函数的逼近是有偏差的。纵然有相应的工作研究不定核函数或者非平移不变核函数的随机特征逼近问题，然而目前不定核函数或者非平移不变核函数的随机特征存在有偏或者方差较大的问题。针对不定核函数或者非平移不变核函数设计无偏且逼近方差较小的随机特征是本文研究的其中一个挑战。

2) **随机特征进一步压缩**

目前核低维随机特征的工作主要围绕着通过减小随机特征的逼近误差，从而实现在相同误差情况下采用更少的低维随机特征数目进行逼近来逼近核函数。然而在提供的计算存储资源有限的情况下，这时候需要进一步减少核低维随机特征的数目来满足计算存储的要求。然而采样值矩阵的秩与最终核方法的泛化性能之间正相关，继续减少核低维随机特征数目将会损害核方法在大数据集上的泛化性能[86, 87, 88]。本文重点研究了在存储资源有限的情况下如何更好地构建核函数的随机特征，从而达到模型性能和体积之间更好的平衡。

3) **从随机特征角度解释神经网络训练过程出现的现象**

深度神经网络成为如今机器学习领域最热门的方法，具有广阔的落地

19004035

应用领域，然而深度神经网络的“黑盒”性质限制其在可靠性要求较高的领域的应用。随机特征和深度神经网络在特征提取方面具有相似性，研究者们尝试从核函数和随机特征的角度解释深度神经网络的收敛性能和泛化性能。一些研究从随机特征的角度很好地分析了深度神经网络在过参数化的时候出现的经验风险和期望风险双下降的现象[33, 34, 89]。此外，在神经网络无限宽的情况下，深度神经网络的动态特性和神经正切核（NTK: Neural Tangent Kernel）[90, 91] 相关，相关研究通过随机特征的方法逼近神经正切核[92]，从而更好地研究深度神经网络的动态特性。最近一些研究关注了神经网络剪枝和随机特征之间的关系，Malach 等人[93] 指出一个随机初始化并且过参数化的网络拥有一个性能相同但参数更少的子网络，在他们分析中网络剪枝相当于构建随机特征模型。深度神经网络的剪枝说明神经网络的优化过程是在较低的维度进行的，本文也将从核函数的低维逼近的角度对深度神经网络的低维优化过程及其潜在的应用进行研究。

1.3 本文主要内容

在前文中，本文回顾了核函数低维随机特征领域的研究现状并且指出目前核函数低维随机特征领域关注的主要问题。本文基于这些问题，研究了随机特征在更广泛的核函数空间与更加严苛的算法落地环境下的构建方法，并且从核函数低维结构的角度探讨了深度神经网络低维训练及其在小样本学习问题上潜在的应用。本文的研究贡献总结如下：

- **提出针对不定核函数和非平移不变核函数的随机傅里叶特征逼近方法**

  针对随机傅里叶特征对核函数的正定性和非平移不变性两个要求，本文提出了在复数空间下构建不定核函数的随机傅里叶特征，突破了随机傅里叶特征对于核函数空间的约束。本文理论证明不定核函数的随机傅里叶特征的无偏性，并且给出了逼近方差减小的方法，实现了针对不定核函数和非平移不变核函数的无偏且低方差的逼近。最后实验验证了此逼近方法相比现有的不定核与非平移不变核逼近方法能够取得更好的核函数逼近效果和性能。

- **提出在计算存储资源有限的情况下随机特征的低比特量化构建方法**

19004035

针对在计算存储资源有限的情况下随机特征维度的减少带来核方法泛化性能下降的问题，本文提出了在计算存储资源有限的情况下构建低比特量化的随机特征，以及相应的基于低比特量化的随机特征的深层核学习框架。相比于 32 位浮点数表示，低比特量化的随机特征允许深层核函数在计算存储资源限制的情况下宽度和深度的增加。本文理论分析低比特量化能够带来的模型压缩和运算加速的程度。最后在不同的计算存储资源下本文验证了在存储资源有限的情况下随机特征的低比特轻量化表示方法比 32 位浮点数表示方法更好的学习性能。

- **基于 NTK 低维假设构造了深度神经网络在小样本学习问题上的优化空间**

  针对深度神经网络在原参数空间与小规模数据集优化过程中出现的“过拟合”情况，本文基于 NTK 低维假设和神经网络低维训练特性，构造了深度神经网络在小样本学习问题上的优化空间，验证了其在小样本学习问题上较优的性能。基于此优化空间，本文提出通过元学习的方法对优化空间进行学习和调整。具体实验上，在基于 NTK 低维假设和元学习方法学习得到的优化空间训练后，深度神经网络在小样本学习问题上出现的“过拟合”情况能够有效缓解。

1.4 本文的组织结构

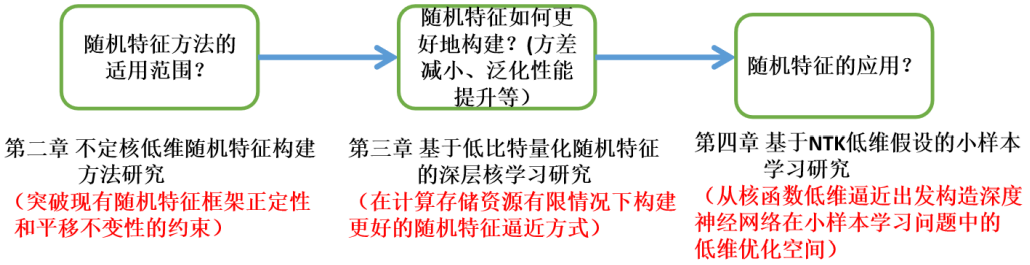


图 1-5　本文章节的组织结构

Fig. 1-5　Organizational structure in this thesis

本文围绕核函数低维随机特征这个主题，旨在解决核函数随机特征构建和应用中的若干问题，图1-5是本文章节的组织结构。本文的章节组织结

19004035

构按照的是现有核函数随机特征研究的整体思路，从随机特征的适用范围、随机特征的构建方法以及随机特征在其他机器学习研究及实际落地场景中的应用三个角度展开。

论文的第一章阐述了核函数随机特征研究的背景与意义，为了读者更好地理解后续章节，此章节介绍了核方法的基础理论。此章节简要回顾了现有核函数随机特征研究的相关工作，总结现有研究的总体脉络及存在问题。根据现有研究的问题，此章节最后介绍了本文的主要内容、研究贡献以及本文各章节的组织结构。

论文的第二章旨在解决现有随机傅里叶特征逼近框架对核函数正定性和平移不变性的约束，提出从复数空间角度逼近不定核及平移不变核函数，理论和实验上证明了提出的逼近方法具有良好的逼近性能。

论文的第三章关注在计算存储资源有限的情况下核随机特征逼近的问题，提出了低比特量化随机特征及其相应的深层核学习框架，在不同计算存储限制下验证了基于低比特量化随机特征的深层核学习框架的性能。

论文的第四章基于神经正切核的低维逼近特性，构造了深度神经网络在小样本学习问题中的低维优化空间，并且在小样本分类问题上进行了实验验证。本章旨在从核函数低维逼近的角度解决深度神经网络在原参数空间和小规模数据集上优化出现的“过拟合”问题。

最后，论文的第五章总结全文，并就本文研究存在的局限性和未来可能的研究方向进行初步探讨。

19004035

第二章　不定核低维随机特征构建方法研究

低维逼近是机器学习模型低维结构研究的重要方面，随机傅里叶特征是核方法领域中低维逼近研究的代表。传统的核方法的运算和空间复杂度依赖于数据样本数目，在大规模核学习问题上传统核方法的运算和空间复杂度较大，难以进行拓展。2006 年，Ali Rahimi 和 Benjamin Recht 提出了随机傅里叶特征方法[22] 来解决大规模核学习中运算与空间复杂度问题。凭借其无偏性以及较小的逼近方差，随机傅里叶特征成为了核学习领域最重要的方法之一。然而随机傅里叶特征需要核函数满足两个性质，平移不变性以及正定性。而部分核函数不满足这两个性质，如常用的线性核、多项式核等。因此本章从正定且平移不变核函数的随机傅里叶特征出发，分析对于不定核及非平移不变核直接应用随机傅里叶特征带来的问题，根据问题提出了在复空间下对于不定核的随机特征逼近的构建方法，并给出了理论分析与实验性能验证。

2.1　理论背景

2.1.1　随机傅里叶特征逼近

定义 2.1. （**测度与** Borel **测度**[94]）假设 X 是一个集合，$\mathcal{A}$ 是定义在 X 的子集上的 σ-代数，则一个函数 $\mu:\mathcal{A}\to[0,+\infty]$ 满足以下两个性质：

1) $\mu(\Phi)=0$

2) $\mu\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\mu\left(A_i\right)$

其中 $\{A_i\}$ 是属于 $\mathcal{A}$ 的无穷不相交序列，性质 2) 也可以称为 σ-可加性。那么则可以称 μ 是定义在 $\mathcal{A}$ 上的测度。若集合 X 是实向量集合 $\mathbb{R}^d$，则 μ 为 Borel 测度。

上面给出了测度和 Borel 测度的定义，当 $\mu(X)<+\infty$，则 μ 是有限测度。当 $\mu(X)=1$ 时则称 μ 是概率测度并且三元组 $(X,\mathcal{A},\mu)$ 是对应的概率空间。$||\mu||$ 称作测度 μ 的总质量，假设 $x\in X$，则计算如下：

$$||\mu||=\int_X|\mu(x)|dx \tag{2.1}$$

19004035

核函数的随机傅里叶逼近理论基础来源于 Bochner 定理，这个定理主要从傅里叶变换与测度论的角度给出了正定核函数的重要性质。

定理 2.1. （Bochner **定理**[95]）定义连续且平移不变核函数 $k(.,.):\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$，核函数 $k(.,.)$ 是正定核函数的充要条件是其可以表示成如下形式：

$$k(x-y)=\int_{\mathbb{R}^d}\exp\left(\mathrm{i}\omega^{\top}(x-y)\right)p(\mathrm{d}\omega) \tag{2.2}$$

其中 $p(\omega)$ 是定义在频率 ω 上的非负且有限的 Borel 测度。

在 Bohner 定理中，i 是虚数符号。如果核函数 $k(x-y)$ 在零向量点处的值为 1，那么根据 Bohner 定理有

$$\int_{\mathbb{R}^d}p(\mathrm{d}\omega)=k(0)=1 \tag{2.3}$$

这说明此时 $p(\omega)$ 是概率测度，那么可以采用 Monte-Carlo 采样方法求取公式(2.2)中的积分，即：

$$\begin{aligned}k(x-y)&=\mathbb{E}_{\omega\sim p(\omega)}\left[\exp(\mathrm{i}\omega^{\top}(x-y))\right]\\&=\mathbb{E}_{\omega\sim p(\omega)}\left[\cos(\omega^{\top}(x-y))\right]+\mathrm{i}\mathbb{E}_{\omega\sim p(\omega)}\left[\sin(\omega^{\top}(x-y))\right]\end{aligned} \tag{2.4}$$

由于核函数 $k(x-y)$ 是实函数，则虚数部分为 0。假设从概率测度 $p(\omega)$ 抽样的频率为 $\omega_1,\omega_2...\omega_s$，则

$$\begin{aligned}k(x-y)&=\mathbb{E}_{\omega\sim p(\omega)}\left[\cos(\omega^{\top}(x-y))\right]\\&\approx\frac{1}{s}\sum_{j=1}^{s}[\cos(\omega_{\mathrm{j}}^{\mathrm{T}}(x-y))]\end{aligned} \tag{2.5}$$

定义关于样本 x 的映射函数 $\psi(x)$ 如下：

$$\psi(x)=\frac{1}{\sqrt{s}}\left[\cos\left(\omega_1^{\mathrm{T}}x\right),\ldots,\cos\left(\omega_{\mathrm{s}}^{\mathrm{T}}x\right),\sin\left(\omega_1^{\mathrm{T}}x\right),\ldots,\sin\left(\omega_s^{\mathrm{T}}x\right)\right]^T \tag{2.6}$$

则公式(2.5)可以写成：

$$k(x-y)\approx\langle\psi(x),\psi(y)\rangle \tag{2.7}$$

19004035

$\psi(x)$ 是核函数 k 的随机傅里叶特征，核函数 k 可以表示成两个随机傅里叶特征的内积。令 Z 为数据集 X 所有样本的随机傅里叶特征的集合，即

$$Z = [\psi(x_1), \psi(x_2), ..., \psi(x_N)] \tag{2.8}$$

则核矩阵 K 可以表示成：

$$K \approx Z^T Z \tag{2.9}$$

由于核函数随机傅里叶特征的维度是 $2s$，其中 s 的维度远小于 N，SVM、SVR 等核方法可以从原空间中进行求解，时间复杂度从 $\mathcal{O}(N^2 d)$ 降低到 $\mathcal{O}(Ns^2)$，空间复杂度从 $\mathcal{O}(N^2)$ 降低到 $\mathcal{O}(Ns)$。核函数随机傅里叶特征从低维角度实现对于高维甚至无穷维的核函数的逼近，$N \times s$ 的随机傅里叶特征矩阵实现了对于 $N \times N$ 的核矩阵的低秩逼近。

2.1.2 不定核与非平移不变核

随机傅里叶特征的构建需要核函数满足两个性质，一个是具有平移不变性，另一个要求是核函数是正定的。核函数广义上是样本之间的相似性度量，很多相似性度量能够在某些实际应用任务中取得很好的效果，却不具有正定性或平移不变性。本小节主要是对不定核函数的相关概念与应用中常见的不定核函数进行介绍，并且给出了在数据归一化在单位球面的情况下，非平移不变核可以转化成平移不变但非正定的核函数。

定义 2.2. (Krein **空间**[96]) *一个内积空间 $\mathcal{H}_k$ 是 Krein 空间的充要条件是存在两个希尔伯特空间 $\mathcal{H}_+$ 和 $\mathcal{H}_-$ 使得*

1）*对于任意的 $f \in \mathcal{H}_k$，有 $f = f_+ \oplus f_-$，其中 $f_+ \in \mathcal{H}_+$，$f_- \in \mathcal{H}_-$。*

2）*对于任意的 $f, g \in k$，$\langle f, g \rangle_{\mathcal{H}_k} = \langle f_+, g_+ \rangle_{H_+} - \langle f_-, g_- \rangle_{H_-}$。*

根据对于 Krein 空间的空间的定义可以看出，Krein 空间能够分解成两个希尔伯特空间的直和的形式。接下来本节进一步定义在核函数上的 Krein 空间。假设 $\mathcal{X}$ 是定义域，$\mathbb{R}^{\mathcal{X}}$ 是定义在 $\mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ 的函数集合，则评价函数的定义如下：

定义 2.3. （**评价函数**[96]） *T_x: $\mathcal{H} \to \mathbb{R}$, 其中 $f \mapsto T_x f = f(x)$。*

19004035

基于定义2.2和2.3，本小节定义再生核 Krein 空间（RKKS: Reproduce Kernel Krein Space），即：

定义 2.4. （**再生核 Kerin 空间**[96]）一个 Krein 空间 $\mathcal{H}_k$ 是一个再生核 Krein 空间的条件是，$\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$ 且在具有强拓扑性质的 $\mathcal{H}_k$ 空间的评价函数是连续的。

和 RKHS 空间类似，一个 RKKS 空间 $\mathcal{H}_k$ 对应着唯一对称实值函数 $k(x,y)$，其中 $k(x,.) \in \mathcal{H}_k$ 且对于任意的 $f \in \mathcal{H}_k$，有 $\langle f, k(x,.)\rangle_{\mathcal{H}_k} = f(x)$。RKHS 空间对应的是正定核函数，而 RKKS 空间则是对应部分的不定核函数及所有的正定核函数，RKHS 空间是 RKKS 空间的一个子集。在 RKKS 理论体系中，正定分解具有重要意义，其构建起部分的不定核函数与正定核函数之间的联系。

定理 2.2. （**正定分解**[96]）假设核函数 k 是一个 $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 的实值核函数，与核函数 k 关联的 RKKS 空间 $\mathcal{H}_k$ 存在的充要条件是存在正定分解

$$k = k_+ - k_- \tag{2.10}$$

其中 k_+ 和 k_- 是两个 RKHS 空间对应的正定核函数。

定理2.2给出了核函数 k 的正定分解的判断条件，正定分解对核函数 k 并非唯一的，同时并非所有的不定核函数具有正定分解。关于不定核的正定分解的存在性问题是核方法研究领域待解决的问题，由于定理2.2要求对于核函数的正定性进行判断，而判断核函数正定性是困难，因此从 RKKS 角度去验证不定核的正定分解的存在性将带来很大的困难。后文将从随机傅里叶特征的角度给出更容易验证的不定核正定分解存在的充要条件。

与正定核函数对应的 RKHS 空间具有的表示定理类似，RKKS 空间也有类似的表示定理，保证了正定核方法能够类似地延伸到不定核函数领域。针对 RKKS 空间的表示定理如下：

定理 2.3. （RKKS **空间的表示定理**[97]）假设 $\mathcal{H}_k$ 是核函数 k 对应的 RKKS 空间，评价函数 $L\{f, \mathcal{X}\}$ 是连续凸函数，其取决于 $f \in k$，其中对于 $x_i \in \mathcal{X}$，有评估值 $f(x_i)$。设 $\Omega\langle\langle f, f\rangle\rangle$ 是一个具有连续特性的稳定器，其中

19004035

Ω 是一个严格单调的函数。令 $C\{f,X\}$ 是对于函数 f 的系列约束条件，那优化问题可以表述成下面形式：

$$\begin{aligned} \underset{f\in K}{\text{stabilize}} \quad & L\{f,X\}+\Omega\left(\langle f,f\rangle_{\mathcal{H}_k}\right) \\ s.t. \quad & C\{f,X\}\leq d \end{aligned} \tag{2.11}$$

则对于此优化问题可以得到鞍点 f^*，其可以表示成：

$$f^*(x)=\sum_{i=1}^{N}\alpha_i k(x,x_i) \tag{2.12}$$

研究者对不定核学习在 SVM、KRR 以及 PCA 问题中的应用有很深入的探索，在这些优化问题中直接使用不定核函数会使得优化问题的目标函数非凸，从而导致优化问题不能够达到全局最优解。针对这个问题，现有文献中主要是利用正定核矩阵 $\widetilde{K}$ 去逼近不定核矩阵 K，通常的做法是去调整不定核矩阵 K 的特征值分布，如将负数特征值置为 0[99] 及将特征值的符号进行翻转[98] 等。Huang 等人[100] 和 Liu 等人[101] 通过将非正定的核矩阵拆分成两个正定核矩阵 $K=K_+-K_-$，利用凹凸过程（Convex-Concave Procedure）等进行求解。这些方法的目的使得不定核方法的优化过程最终能够接近全局最优解。

下面介绍下常用的不定核函数，常用的不定核函数大体上可以分为四种类型：

- Epanechnikov **核函数**[97]

$$k(x,y)=\left(1-\frac{\|x-y\|_2^2}{a^2}\right)^m \tag{2.13}$$

其中 $||x-y||_2^2\leq a^2$。对于如线性核、多项式核等非平移不变核函数，当样本输入归一化到单位球面上，即 $||x||_2=||y||_2=1$ 的时候，其可以转化成 Epanechnikov 核函数，其中 $||x-y||_2^2=2-\langle x,y\rangle\in[0,2]$。对应

19004035

的转化过程如下：

$$k(x,y)=\left(1-\frac{\|x-y\|_2^2}{a^2}\right)^m=\alpha(q+\langle x,y\rangle)^m \tag{2.14}$$

其中 $q=a^2/2-1$，$\alpha=(2/a^2)^m$，$m\geq 1$，$a\geq 2$。Epanechnikov 核函数的傅里叶变换 $p(\omega)$ 为

$$p(\omega)=\sum_{i=0}^{p}\frac{m!}{(m-i)!}(1-\frac{4}{a^2})^{m-i}(\frac{2}{a^2})^i(\frac{2}{||\omega||_2})^{\frac{d}{2}+i}J_{\frac{d}{2}+i}(2||w||_2) \tag{2.15}$$

- **条件正定核函数**（Conditionally Positive Definite Kernel）

  条件正定核函数指的是在某些超参数的设置下核函数是正定的，而核函数的超参数没有任何限制时，无法确保核函数的正定性。如 tanh 核函数，

  $$k(x,y)=\tanh(\mathrm{c}x^{\mathrm{T}}y+\mathrm{d}) \tag{2.16}$$

  如果 c 和 d 都没有任何范围限制的话可以认为是不定核。又如 Huang 等人提出的 TL1 核函数（Trunacated ℓ_1 distance kernel）[102]：

  $$k(x,y)=\max\{\rho-||x-y||_1,0\} \tag{2.17}$$

  其中 ρ 通常采用 $0.7d$，其中 d 是样本维度。

- **正定核函数的线性组合**

  绪论中提到当正定核函数的线性组合系数为正数，则生成的核函数为正定核函数。若组合系数的取值范围为实数域，则生成的核函数不能保证正定性质。例如高斯核函数的线性组合（Delta-Gaussian Kernel）[97]：

  $$k(x-y)=\sum_{i=1}^{m}a_i\exp\left(-\frac{||x-y||^2}{2\sigma_i^2}\right) \tag{2.18}$$

19004035

其中 $a_i \in \mathbb{R}$ 且 $\sigma_i > 0$，其对应的傅里叶变换 $p(\omega)$ 为

$$p(w) = \sum_{i=1}^{m} a_i \sigma_i^d e^{-\frac{\sigma_i^2 ||w||^2}{2}} \tag{2.19}$$

- **其他的相似性度量**

  核函数广义上代表了样本与样本之间的相似度或者不相似度，其中大部分相似度度量不满足正定性质，如用于字符串匹配和通信编码的核函数[103]、用于衡量图的匹配程度的图核函数[104] 等都是不定核函数。

上面不定核函数中最值得注意的是 Epanechnikov 核函数，因为其为利用随机傅里叶特征方法对非平移不变核函数的逼近提供了方法，即可以通过对数据进行归一化，将所有数据约束在单位球面上，然后非平移不变核函数可以转化成 Epanechnikov 核函数的形式。因此，对于不定核和非平移不变核函数的随机傅里叶逼近可以统一成对于平移不变但非正定的核函数的随机傅里叶特征逼近。

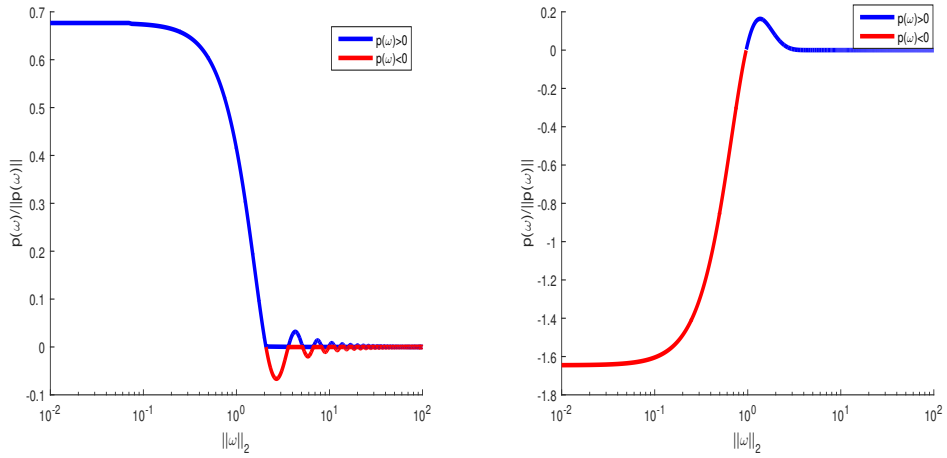
2.2 复空间下的不定核随机特征逼近构建方法

根据 Bohner 定理，正定且平移不变核函数在频域上对应的是非负的 Borel 测度。相应地，对不定核而言，频谱 $p(\omega)$ 不是非正的测度，甚至不是 Borel 测度。图2-1展示的是 $d = 2$ 时 Epanechnikov 核函数和 Delta-Gaussian 核函数的频谱 $p(\omega)$，其可以分为两个部分，蓝色曲线代表 $p(\omega) \geq 0$，红色的曲线代表 $p(\omega) < 0$。非正的测度使得 Monte-Carlo 采样方法不能依据频谱 $p(\omega)$ 采集权值。针对此问题，本小节从符号测度和 Jordan 分解的角度出发，提出了对于不定核的频谱 $p(\omega)$ 的处理策略与复数空间下对于不定核函数的随机傅里叶特征的构建方法，并且从不定核的随机傅里叶特征的角度给出了更容易验证的不定核正定分解存在性的条件。

2.2.1 不定核的随机傅里叶特征

定义 2.5. （**符号测度**[94]）令 X 是一个集合，$\mathcal{A}$ 是定义在 X 的子集上的 σ-代数，函数 $\mu : \mathcal{A} \to [-\infty, +\infty)$ 或者 $\mathcal{A} \to (-\infty, +\infty]$ 满足 σ-可加性，则函数 μ 是符号测度。

19004035



(a) Epanechnikov 核函数 ($a=3, m=1$)　(b) Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=2$)

图 2-1　Epanechnikov 核函数和 Delta-Gaussian 核函数的频谱 $p(\omega)$

Fig. 2-1　Spectrum of Epanechnikov kernel function and Delta-Gaussian kernel function $p(\omega)$

和测度相比，符号测度允许值域为负数，是测度这个概念的延伸。基于符号测度的概念，下面给出的 Jordan 分解说明符号测度可以分解成两个非负的测度。

定理 2.4.　(Jordan **分解**[94]) 令 μ 是定义在 σ-代数 $\mathcal{A}$ 上的符号测度，存在两个非负的测度（其中一个测度是有限的）μ_+ 和 μ_-，使得符号测度 $\mu=\mu_+-\mu_-$。

值得注意的是，Jordan 分解并非唯一。与 Jordan 分解相关的 Hahn 分解定理阐述了集合 X 可以分解为两部分，即 $X=X_+\cup X_-$。其中对于集合 X_+ 上的任意子集 S_+ 都有 $\mu(S_+)\geq 0$，对于集合 X_- 上的任意子集 S_- 都有 $\mu(S_-)\leq 0$。符号测度的总质量（Total Mass）$||\mu||$ 是两个非负测度总质量的和，即 $||\mu||=||\mu_+||+||\mu_-||$。

不定核的傅里叶变换 $p(\omega)$ 不是非负测度，但可以从符号测度的角度进行思考，利用 Jordan 分解将 $p(\omega)$ 拆分成两个非负测度。假设 $p(\omega)$ 对应

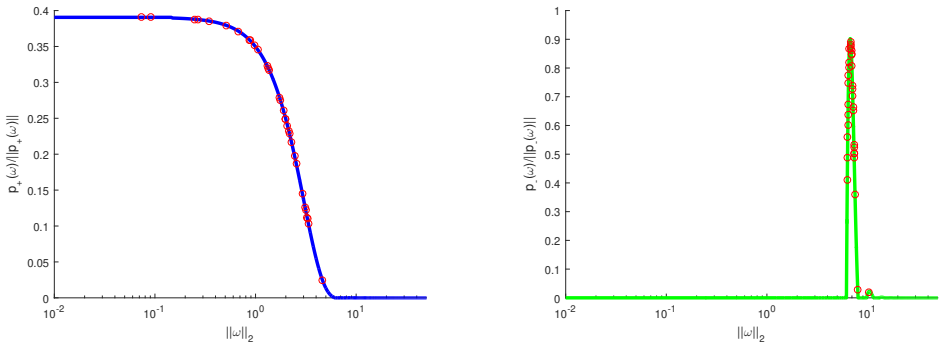
19004035

的两个非负测度为 $p_+(\omega)$ 和 $p_-(\omega)$，则

$$\begin{aligned} k(x-y) &= \int_{\mathbb{R}^d} \exp(\mathrm{i}\omega^T(x-y))p(d\omega) \\ &= \int_{\mathbb{R}^d} \exp(\mathrm{i}\omega^T(x-y))p_+(d\omega) - \int_{\mathbb{R}^d} \exp(\mathrm{i}\upsilon^T(x-y))p_-(d\upsilon) \\ &= ||p_+||\mathbb{E}_{\omega\sim\widetilde{p}_+(\omega)}\left[\exp(\mathrm{i}\omega^T(x-y))\right] - ||p_-||\mathbb{E}_{\upsilon\sim\widetilde{p}_-(\upsilon)}\left[\exp(\mathrm{i}\upsilon^T(x-y)\right], \end{aligned} \tag{2.20}$$

其中 $||p_+||$ 和 $||p_-||$ 是非负测度 $p_+(\omega)$ 和 $p_-(\upsilon)$ 的总质量，$\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 是归一化后的概率测度，即：

$$\widetilde{p}_+(\omega) = \frac{p_+(\omega)}{||p_+||}, \quad \widetilde{p}_-(\upsilon) = \frac{p_-(\upsilon)}{||p_-||} \tag{2.21}$$



(a) $p_+(\omega)/||p_+(\omega)||$ (b) $p_-(\omega)/||p_-(\omega)||$

图 2-2 Jordan 分解后归一化测度 $\frac{p_+(\omega)}{||p_+||}$ 和 $\frac{p_-(\omega)}{||p_-||}$ 及 Monte-Carlo 采样点

Fig. 2-2 Normalized measure after Jordan decomposition $\frac{p_+(\omega)}{||p_+||}$ and $\frac{p_-(\omega)}{||p_-||}$ as well as Monte-Carlo sampling points

通过 Jordan 分解，本章可以利用 Monte Carlo 的方法对 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 进行采样。图2-2是对 Epanechnikov 核函数的频谱进行 Jordan 分解的结果，其中红色散点是对分解后的测度进行 Monte Carlo 采样得到的点。注意的是本章后续实验中研究的不定核函数都假设具有径向性质，则对应的傅里叶变换也是径向核函数，在这种情况下 $||p_+||$ 和 $||p_-||$ 的计算公式

19004035

如下：

$$||p_+|| = \int_0^\infty |p_+(||\omega||)|d||\omega||, \quad ||p_-|| = \int_0^\infty |p_-(||\upsilon||)|d||\upsilon|| \tag{2.22}$$

根据公式(2.4)，公式(2.20)中的两个期望可以替换成两个正定核函数 $\widetilde{k}_+(x-y)$ 和 $\widetilde{k}_-(x-y)$，可以得到

$$\begin{aligned} k(x-y) &= ||p_+||\widetilde{k}_+(x-y) - ||p_-||\widetilde{k}_-(x-y) \\ &= k_+(x-y) - k_-(x-y) \end{aligned} \tag{2.23}$$

公式(2.23)似乎可以与 RKKS 核函数的正定分解联系起来，但上一小节提到并非所有的不定核具有正定分解。下一小节将从公式(2.23)出发，给出不定核函数的正定分解存在性条件。

假设从 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 分别进行独立同分布 Monte-Carlo 采样，得到的权值为 $\{\omega_i\}_{i=1}^s$ and $\{\upsilon_i\}_{i=1}^s$，则定义关于样本 x 的函数映射 $\phi_i(x)$ 如下：

$$\phi_i(x) = [\sqrt{||p_+||}\cos(\omega_i^T x), \sqrt{||p_+||}\sin(\omega_i^T x), \mathrm{i}\sqrt{||p_-||}\cos(\upsilon_i^T x), \mathrm{i}\sqrt{||p_-||}\sin(\upsilon_i^T x)] \tag{2.24}$$

基于 $\phi_i(x)$，不定核的随机傅里叶特征 $\psi(x)$ 如下：

$$\psi(x) = \frac{1}{\sqrt{s}}[\phi_1(x), \phi_2(x), ..., \phi_s(x)]^T \tag{2.25}$$

类似正定核函数，不定核通过两个样本随机傅里叶特征 $\psi(x)$ 的内积进行低维逼近。

$$k(x-y) \approx \frac{1}{s}\sum_{i=1}^{s}\langle\phi_i(x), \phi_i(y)\rangle = \langle\psi(x), \psi(y)\rangle \tag{2.26}$$

图2-3总结了复空间下的不定核随机特征逼近构建总体的流程。不定核随机特征与正定核随机特征构建上的主要区别在于：

1）需要将 $p(\omega)$ 进行测度分解，并从两个概率分布进行采样。

2）需要在复数空间进行构建。

可以看出当 $||p_-||$ 为零时，公式(2.25)是正定且平移不变核函数的随

19004035

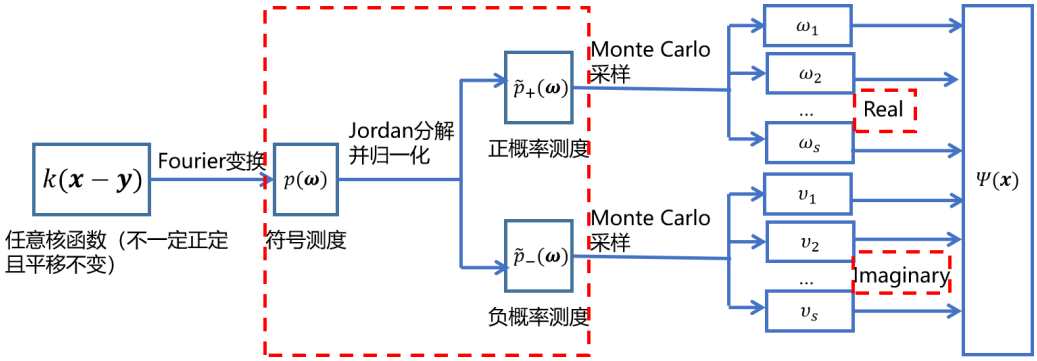


图 2-3　复空间下的不定核随机特征逼近构建的流程框图

Fig. 2-3　The flow chart of the construction of random features of indefinite kernel in complex space

机傅里叶特征。而上一节提到对于不定核与非平移不变核函数的随机傅里叶逼近可以统一成对于平移不变但非正定的核函数的随机傅里叶特征逼近。因此可以称公式(2.25)是广义构建的核随机傅里叶特征（GRFF: Generalized Random Fourier Features）。类似地，GRFF 构建方法可以将核方法的运算和存储复杂度下降到 $\mathcal{O}(4Ns)$ 和 $\mathcal{O}(4Ns^2)$。

2.2.2　不定核正定分解的存在性条件

回顾定理2.2，其阐明的一个核心的问题是**给定一个不定核函数，是否存在对应的正定分解？**非正定的核矩阵可以通过特征值分解的方式拆分成两个正定核矩阵之差，而不定核函数则不一定能够拆分成两个正定核函数的差。命题2.2给出的 RKKS 空间判定条件只适用于部分容易判定的不定核，如组合系数为负数的正定核函数的线性组合，对于大部分的不定核这个准则难以验证。利用不定核的随机傅里叶特征，可以给出更容易进行验证的不定核正定分解的存在性条件。

定理 2.5.（**不定核正定分解的存在性定理**）给定不定核函数 k，其（广义）傅里叶变换是 $p(\omega)$。不定核函数 k 有正定分解 $k=k_+-k_-$ 的充分必要条件是 $p(\omega)$ 的总质量是有限的，即 $||p||<+\infty$。

定理2.5的证明由附录 A 给出。根据定理2.5，通过求取不定核函数的傅

19004035

里叶变换就可以进行不定核函数正定分解存在性的验证，绕开了对于核函数的正定性判断。

2.3 无偏性证明与方差减小策略

无偏性和有效性是核函数的低维随机特征逼近方法重点考察的两个维度，对应的是逼近的偏差和方差。本小节首先分别给出无偏性和有效性的定义。

定义 2.6. （**无偏性**[105]）设 $X_1, X_2, ..., X_s$ 是随机变量，参数 θ 的估计量为 $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_s)$，记随机变量的期望为符号 $\mathbb{E}(.)$，若有

$$\mathbb{E}(\hat{\theta}) = \theta \tag{2.27}$$

则称 $\hat{\theta}$ 是 θ 的无偏估计。如 $\mathbb{E}(\hat{\theta}) \neq \theta$，则称 $\varepsilon = \mathbb{E}(\hat{\theta}) - \theta$ 为估计量 θ 的偏差。

定义 2.7. （**有效性**[105]）设 $\hat{\theta}_0$ 是未知参数 θ 的一个无偏估计量，随机变量的方差为符号 $\mathbb{V}(.)$。如果在所有 θ 的无偏估计量 $\hat{\theta}$ 中均有

$$\mathbb{V}(\hat{\theta}_0) \leq \mathbb{V}(\hat{\theta}) \tag{2.28}$$

成立，则称 $\hat{\theta}_0$ 是 θ 的有效估计量。

无偏性是核函数的随机特征逼近最基本的性质，假设利用 GRFF 内积计算得到的核函数近似值为 $K_{\mathrm{GRFF}}(z)$，即

$$K_{\mathrm{GRFF}}(z) = \langle \psi(x), \psi(y) \rangle \tag{2.29}$$

其中 $z = x - y$，$\psi(.)$ 是构建的 GRFF。下面的定理给出了 GRFF 方法的无偏性质。

定理 2.6. $K_{\mathrm{GRFF}}(z)$ 是对于所有的平移不变核函数 $k(z)$ 的无偏估计，即

$$\mathbb{E}[K_{\mathrm{GRFF}}(z)] = k(z) \tag{2.30}$$

19004035

逼近的方差是：

$$\mathbb{V}[K_{\mathrm{GRFF}}(z)] = \frac{||p_+||^2}{s}\left[\frac{1+\widetilde{k}_+(2z)}{2} - \widetilde{k}_+^2(z)\right] + \frac{||p_-||^2}{s}\left[\frac{1+\widetilde{k}_-(2z)}{2} - \widetilde{k}_-^2(z)\right] \tag{2.31}$$

定理2.6的证明由附录 B 给出，此定理说明逼近方差决定了 GRFF 最终的逼近误差。类似 RFF 和 ORF 的关系，我们可以采用对于采样值矩阵进行正交化来减小 GRFF 的逼近误差，具体的步骤如下：

1) **采样随机权值的模**

分别从 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 采样随机权值 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$ 的 ℓ_2 范数的模，即

$$||\omega_i||_2 \sim \widetilde{p}_+(||\omega||), \quad ||\upsilon_i||_2 \sim \widetilde{p}_-(||\upsilon||) \tag{2.32}$$

2) **采样随机权值对应的方向向量并对各方向向量正交化**

假设从标准正态分布中采样的方向向量为 $\{a_i\}_{i=1}^m$ and $\{b_i\}_{i=1}^m$，其中 $m = \max(\mathrm{s},\mathrm{d})$。采样的方向向量可以组合成随机矩阵 M，对随机矩阵 M 做 QR 分解得到随机权值的正交方向向量。

$$\begin{aligned} &a_j \sim N(0, I_{2m}), \quad b_j \sim N(0, I_{2m}) \\ &M = [a_1, ..., a_m, b_1, ..., b_m], \quad M^{orth} = \mathrm{QR(M)} \end{aligned} \tag{2.33}$$

3) **权值的模和正交后的方向向量融合**

从正交矩阵 M^{orth} 中取前 d 行并归一化逐列，构成新矩阵 M^{orthn}。采样的模与正交单位方向向量可以构成正交化的随机权值 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$，即

$$\omega_i = ||\omega_i||_2 M_i^{orthn}, \quad \upsilon_i = ||\upsilon_i||_2 M_{s+i}^{orthn} \tag{2.34}$$

进行正交化后，本节构造了逼近误差更小的 GRFF，正交化后的 GRFF 称为广义构建的正交化随机傅里叶特征（GORF：Generalized Orthogonal Random Features）。GORF 的构建方法总结如算法1所示。

与 GRFF 不同，GORF 通过正交化的方法，使得随机权值不再满足独立同分布采样。**这种正交化采样方式能够带来多少逼近方差的下降呢？**在推导针对不定核函数采用正交采样方式前后逼近方差的变化之前，本节先

19004035

给出对于正定核函数采用正交采样前后逼近方差的变化。

算法 1 GORF 构建方法

输入: 平移不变核函数 $k(x,y)=k(z)$, $z=||x-y||$, 训练样本的维度 d, GORF 的数目 s

输出: 满足 $k(x-y)\approx\phi(x)^T\phi(y)$ 的 GORF 映射 $\phi(x)$

1: 通过核函数的随机傅里叶变换获取符号测度 $p(\omega)$, 并且通过 Jordan 分解与归一化计算 $p_+(\omega)$, $p_-(\upsilon)$, $||p_+||$, $||p_-||$, $\widetilde{p}_+(\omega)$, $\widetilde{p}_-(\upsilon)$。

2: 分别从 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 中采样 s 个 ℓ_2 范数的权值的模，即 $||\omega_i||_2$ 和 $||\upsilon_i||_2$。

3: 从标准正态分布 $N(0,I_{2m})$ 采样 $2m$ 方向向量 a_j 和 b_j，并组合成矩阵 M。利用 QR 分解获得正交矩阵 M^{orth} 和正交单位矩阵 M^{orthn}。

4: 根据公式(2.34) 得到 ω_i 和 υ_i。

5: 根据公式(2.24)和(2.25)产生 $\psi_i(x)$ 和 $\phi(x)$。

定理 2.7. [65] 给定在 $\mathbb{R}^d$ 上定义的正定且平移不变核函数 k，其对应的（广义）傅里叶变换为 $p(\omega)$。$K_{\mathrm{RFF}}(.)$ 和 $K_{\mathrm{ORF}}(.)$ 分别是采用 RFF 和 ORF 后对核函数 k 的无偏估计。对于任意 $x,y\in\mathbb{R}^d$，记 $z=x-y$，则有

$$G_k(z)=\mathbb{V}(K_{\mathrm{ORF}}(z))-\mathbb{V}(K_{\mathrm{RFF}}(z))$$
$$=\frac{s-1}{s}\mathbb{E}_{R_1}\left[\frac{J_{\frac{d}{2}-1}(R_1||z||)\Gamma(d/2)}{(R_1||z||/2)^{\frac{d}{2}-1}}\right]^2-\frac{s-1}{s}\mathbb{E}_{R_1,R_2}\left[\frac{J_{\frac{d}{2}-1}(\sqrt{R_1^2+R_2^2}||z||)\Gamma(d/2)}{(\sqrt{R_1^2+R_2^2}||z||/2)^{\frac{d}{2}-1}}\right], \tag{2.35}$$

其中 $R_1,R_2\sim p(\omega)$，并且 J_α 是自由度为 α 的第一类贝塞尔函数。

根据定理2.7，下面给出对于不定核函数采用正交采样方式前后逼近方差的变化。

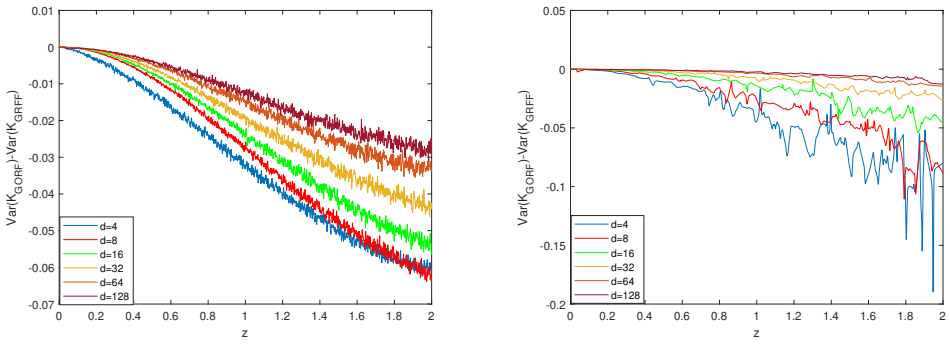
定理 2.8. 给定在 $\mathbb{R}^d$ 上定义的不定核函数 k，其对应的（广义）傅里叶变换为 $p(\omega)$，$||p(\omega)||<+\infty$。$K_{\mathrm{GRFF}}(.)$ 和 $K_{\mathrm{GORF}}(.)$ 分别是采用 GRFF 和 GORF 后对不定核函数 k 的无偏估计。对于任意 $x,y\in\mathbb{R}^d$，记 $z=x-y$，则有

$$\mathbb{V}[K_{\mathrm{GORF}}(z)]-\mathbb{V}[K_{\mathrm{GRFF}}(z)]=||p_+||^2G_{\widetilde{k}_+}(z)+||p_-||^2G_{\widetilde{k}_-}(z)+H(z) \tag{2.36}$$

19004035

其中 $G_{\widetilde{k}_+}(z)$ 和 $G_{\widetilde{k}_-}(z)$ 可以通过定理2.7计算，并且 $H(z)=2||p_+||||p_-||[\mathbb{E}(a_1)\mathbb{E}(b_1)-\mathbb{E}(a_1b_1)]$，$a_1=\cos(\omega_1^T z)$，$b_1=\cos(\upsilon_1^T z)$。

定理2.8的证明由附录 C 给出。定理2.8的公式(2.36)将方差减小量分成两个部分，$||p_+||^2G_{\widetilde{k}_+}(z)+||p_-||^2G_{\widetilde{k}_-}(z)$ 代表的是对于 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$ 分别进行正交化减小的方差，$H(z)$ 代表的是 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$ 相互之间正交减小的方差。当 $||p_+||^2G_{\widetilde{k}_+}(z)+||p_-||^2G_{\widetilde{k}_-}(z)+H(z)\leq 0$ 时，GORF 在取值 z 上比 GRFF 的逼近方差小。



(a) Epanechnikov 核函数 $(a=3,m=1)$　(b) Delta-Gaussian 核函数 $(a_1=1,a_2=-1,\sigma_1=1,\sigma_2=10)$

图 2-4　不同数据维度和不定核函数下数值计算的 $\mathbb{V}[K_{\mathrm{GORF}}(z)]-\mathbb{V}[K_{\mathrm{GRFF}}(z)]$ $(z=||x-y||_2,\ s=d)$

Fig. 2-4　Numerical calculations for $\mathbb{V}[K_{\mathrm{GORF}}(z)]-\mathbb{V}[K_{\mathrm{GRFF}}(z)]$ $(z=||x-y||_2,\ s=d)$ under different data dimensions and indefinite kernel functions

根据定理2.8，对于不同的不定核函数，GORF 比 GRFF 方法的逼近方差的减小量可以通过数值方法计算。本节利用 Monte-Carlo 方法在 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 采样 10000 个随机权值来计算定理2.8中的期望。图2-4展示的是对于 Epanechnikov 核函数和 Delta-Gaussian 核函数采用了 GORF 后减小的逼近方差，图中的横轴是两个样本之间的均方误差距离。从图中可以看出，在较大范围的均方误差取值范围中，$\mathbb{V}[K_{\mathrm{GORF}}(z)]-\mathbb{V}[K_{\mathrm{GRFF}}(z)]\leq 0$，GORF 能够带来显著的逼近方差的减小。对于不同的数据维度，$\frac{\mathbb{V}[K_{\mathrm{GORF}}(z)]-\mathbb{V}[K_{\mathrm{GRFF}}(z)]}{\mathbb{V}[K_{\mathrm{GRFF}}(z)]}$ 保持恒定，说明对于不同维度的数据 GORF 都能够带来显著的逼近方差的减小。

19004035

2.4 实验验证

这节主要对不定核的随机傅里叶逼近 GRFF 和 GORF 进行验证，在逼近误差及分类回归等具体问题的性能方面，对比提出的 GRFF 与 GORF 方法和现有的不定核的低维逼近方法。本节的所有实验硬件环境是带有 i7-4970 CPU（频率 3.6GHz）和 32G 的 RAM 的 PC 机，实验的软件环境是 Matlab。

2.4.1 实验设置

本章主要研究的是针对不定核函数的低维逼近方式,实验采用了 Epanechnikov 核函数与 Delta-Gaussian 核函数，在三个分类数据集与两个回归数据集上进行实验验证。实验采用的核函数的形式与对应的傅里叶变换在 2.2 节给出，其中对于 Epanechnikov 核函数超参数设置为 $a = 3, m = 1$，对于 Delta-Gaussian 核函数超参数设置为 $a_1 = 1, a_2 = -1, \sigma_1 = 1, \sigma_2 = 10$。下面主要介绍实验中所使用的数据集，其中数据集中的输入归一化到 $[0,1]^d$ 范围内。

1）分类任务的数据集

- *letter* 数据集　对于字母进行二分类的 UCI 数据集，训练集总共有 12000 个样本，测试集有 6000 个样本，样本维度为 16。
- *ijcnnl*1 数据集　IJCNN2001 机器学习挑战赛所使用的二分类数据集，训练集总共有 49990 个样本，测试集有 91701 个样本，样本的维度为 22。
- *usps* 数据集　Kaggle 比赛中 16*16 的手写数字分类数据集，训练集总共有 7291 个样本，测试集有 2007 个样本，数据的维度为 256。

2）回归任务的数据集

- *mpg* 数据集　数据集来源于美国 CMU 维护的 StatLib 库，其包含 1999 年和 2008 年 38 款流行车型的燃油经济性数据，预测指标为燃油效率，数据集总共有 398 个样本，样本维度为 7，其中随机选取 319 个样本作为训练集，79 个样本作为测试集。

19004035

- *housing* 数据集　波士顿房价预测的 UCI 数据集，数据集总共有 506 个样本，样本维度为 13。其中随机选取 405 个样本为训练集，选取 101 个样本作为测试集。

2.4.2 不同核低维逼近方法的逼近误差对比

逼近误差（approximation error）是核函数低维随机逼近方法中最重要的评价指标，逼近偏差和方差决定了最终核函数低维随机逼近方法的逼近误差。量化逼近误差的指标是在 1000 个数据采样点上估计的核矩阵 $\hat{K}$ 和实际核矩阵 K 之间的相对误差，即

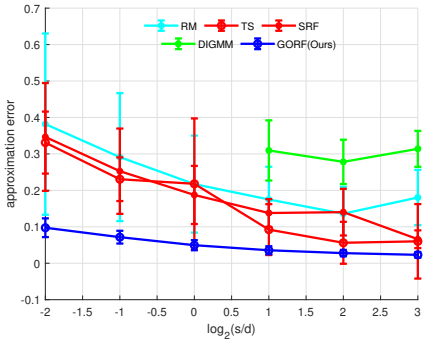
$$approximation\ error = \frac{||K-\hat{K}||_F}{||K||_F} \tag{2.37}$$

其中 $||.||_F$ 是矩阵的 Frobenius 范数。实验中展示的是 10 次重复实验中逼近误差的均值和标准差。

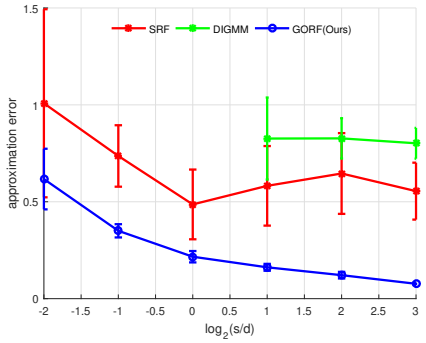
本小节首先对比了 GORF 方法和现有提出的不定核函数或非平移不变核函数的逼近方法。RM[46] 和 TS 方法[47] 是针对多项式核函数的无偏的随机逼近方法，基于的是 Maclaurin 展开。SRF[84] 和 DIGMM[85] 是针对不定核函数提出的随机特征逼近方法，利用高斯混合模型对核函数的傅里叶变换进行逼近，高斯混合模型逼近的方法不是无偏的，有一定逼近偏差。图2-5是在三种分类数据集与两种不定核函数上提出的 GORF 方法与现有随机逼近方法的逼近误差对比实验，横轴是随机特征数目 s 与数据维度 d 的比例的对数值，纵轴是逼近误差。由于计算 PolyGamma 函数的时候出现复数计算，实验中 DIGMM 方法的验证从 $s \geq 2d$ 开始。实验结果显示，相比其他的不定核随机特征逼近方法，本章提出的 GORF 方法在不同的随机特征数目 s 下能够取得最低的逼近误差。GORF 能够取得最低的逼近误差主要取决于两个方面，一个相较于 SRF 和 DIGMM 利用高斯混合模型逼近的方法，GORF 直接对于测度进行了 Jordan 分解，保证了其没有偏差；二个是相较于 RM 和 TS，随机傅里叶特征本身的逼近方差较小，并且在此基础上采用了正交采样方式进一步减小逼近方差。无偏性和较小的方差保证了 GORF 方法较小的逼近误差。

为了进一步验证正交采样方法对于减小逼近方差的作用，表格2-1对比

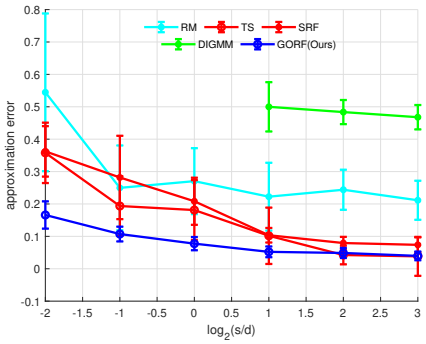
19004035



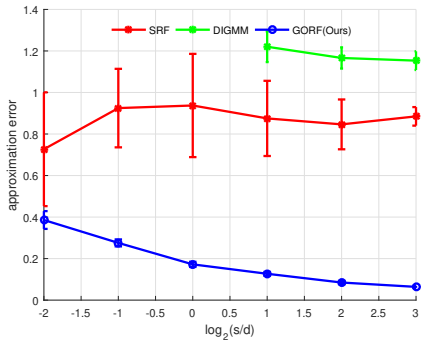
(a) *letter* 数据集，Epanechnikov 核函数 ($a=3, m=1$)



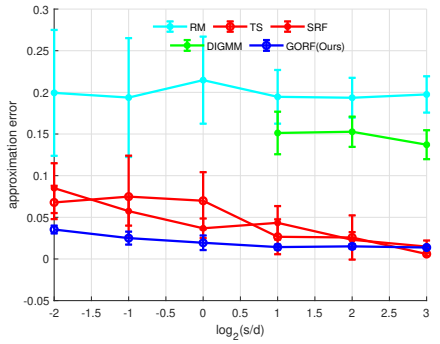
(b) *letter* 数据集，Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=10$)



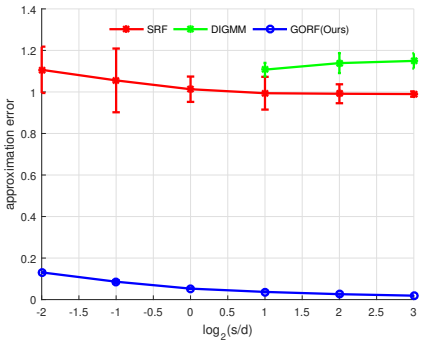
(c) *ijcnn1* 数据集，Epanechnikov 核函数 ($a=3, m=1$)



(d) *ijcnn1* 数据集，Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=10$)



(e) *usps* 数据集，Epanechnikov 核函数 ($a=3, m=1$)



(f) *usps* 数据集，Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=10$)

图 2-5　多种不定核或非平移不变核的随机逼近方法在不同核函数与数据集下逼近误差的对比实验

Fig. 2-5　Contrast experiments on the approximation errors of a variety of random approximation methods for various indefinite kernels or non-translation invariant kernels under different kernel functions and datasets

19004035

表 2-1 GRFF 和 GORF 在不同核函数和数据集上逼近误差对比

Table 2-1 Comparison of approximation errors between GRFF and GORF on different kernel functions and data sets

核函数	数据集	方法	s = 1/2d	s = d	s = 2d
Epanechnikov	*letter*	GRFF	0.0859±0.0309	0.0547±0.0078	0.0469±0.0109
		GORF	**0.0716±0.0175**	**0.0495±0.0139**	**0.0360±0.0110**
	*ijcnn*1	GRFF	0.1159±0.0158	0.0907±0.0194	0.0794±0.0142
		GORF	**0.1072±0.0228**	**0.0775±0.0204**	**0.0487±0.0155**
	usps	GRFF	0.0270±0.0056	0.0213±0.0063	0.0160±0.0029
		GORF	**0.0251±0.0078**	**0.0194±0.0087**	**0.0143±0.0030**
Delta-Gaussian	*letter*	GRFF	0.3918±0.0428	0.2736±0.0345	0.1887±0.0201
		GORF	**0.3154±0.0424**	**0.1133±0.0181**	**0.0760±0.0090**
	*ijcnn*1	GRFF	0.2924±0.0188	0.2171±0.0222	0.1504±0.0134
		GORF	**0.2415±0.0190**	**0.1026±0.0129**	**0.0739±0.0065**
	usps	GRFF	0.1005±0.0061	0.0690±0.0050	0.0500±0.0024
		GORF	**0.0724±0.0049**	**0.0235±0.0009**	**0.0166±0.0008**

GORF 和 GRFF 方法的逼近误差。GRFF 和 GORF 方法都是无偏估计，这两种方法的逼近误差只是由逼近方差决定。从表格2-1中可以看出对于 Epanechnikov 核函数与 Delta-Gaussian 核函数，采用了正交采样的 GORF 方法可以将逼近误差降低到 GRFF 方法的 10% 以上，证明了正交化技巧能够进一步减小不定核随机傅里叶特征的逼近方差。

2.4.3 分类和回归问题上不同逼近方法结果对比

这个部分主要比较了 GORF 方法和其他的低维随机逼近方法在 SVM 分类算法和 SVR 回归算法上的性能。在具体的 SVM 分类和 SVR 回归中，由于各种随机特征逼近方法已经给出了核函数对应的映射函数，因此只需要在原空间通过梯度下降方式求解分类器的参数。求解过程利用了 *liblinear* 工具包[106]，其中工具包超参数 C 是从 [0.01,0.1,1,10,100,1000] 中进行五折交叉验证选择出来的，其他超参数与上一小节保持一致。分类准确率是评价低维随机逼近方法的指标，即分类正确的样本数目占总样本的数目；而在回归任务中，采用的评价指标是均方根误差（RMSE: Rooted Mean Square

19004035

Error)，计算 RMSE 公式如下所示，

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{2.38}$$

其中 $\{y_i\}_{i=1}^{N}$ 和 $\{\hat{y}_i\}_{i=1}^{N}$ 是预测值与对应的真实值。

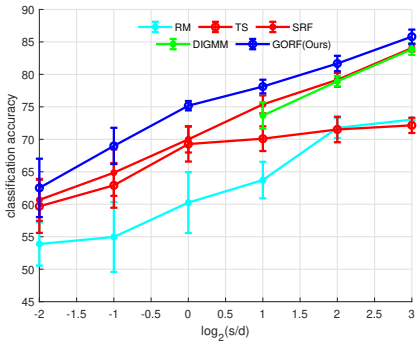
图2-6是多种不定核随机逼近方法在 SVM 分类问题上分类准确率的对比结果，可以看到相对现有的随机逼近方法，GORF 方法能够取得更好的分类准确率。表2-2是多种不定核随机逼近方法在 SVR 回归问题上 RMSE 的对比结果，可以看到 GORF 方法在多个数据集上可以取得更低的 RMSE。同时可以看出随着特征数目的增加，各种不定核函数随机逼近方法的分类准确率提升，并且回归 RMSE 下降。实验结果说明不定核函数随机特征的逼近误差越小，其在分类和回归问题上的性能更好，这符合独立于数据进行采样的随机特征逼近方法的基本假设。

表 2-2　多种随机逼近方法在 SVR 回归问题的 RMSE 对比

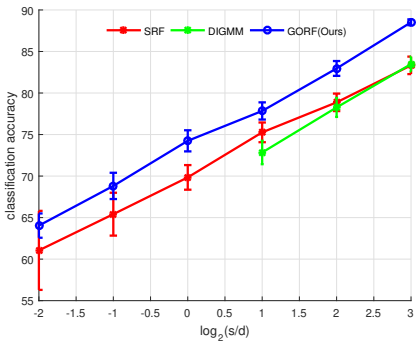
Table 2-2　RMSE comparison of various random approximation methods in SVR regression problem

核函数	数据集	方法	s = 2d	s = 4d	s = 8d
Epanechnikov	*mpg*	RM	7.137 ± 1.828	5.623 ± 0.862	4.707 ± 0.107
		TS	5.050 ± 0.759	4.952 ± 0.423	4.674 ± 0.117
		SRF	4.461 ± 0.136	4.269 ± 0.129	4.137 ± 0.081
		DIGMM	4.686 ± 0.287	4.339 ± 0.128	4.133 ± 0.119
		GORF(OURS)	**4.342 ± 0.102**	**4.162 ± 0.157**	**3.872 ± 0.117**
	housing	RM	7.153 ± 1.772	5.436 ± 0.917	4.491 ± 0.008
		TS	5.414 ± 0.879	4.772 ± 0.377	4.657 ± 0.316
		SRF	4.391 ± 0.368	3.906 ± 0.219	3.555 ± 0.130
		DIGMM	4.897 ± 0.368	4.130 ± 0.324	4.000 ± 0.280
		GORF(OURS)	**4.079 ± 0.233**	**3.817 ± 0.204**	**3.472 ± 0.137**
Delta-Gaussian	*mpg*	SRF	5.243 ± 0.110	5.189 ± 0.095	4.958 ± 0.090
		DIGMM	5.203 ± 0.212	4.925 ± 0.215	4.613 ± 0.111
		GORF(OURS)	**4.831 ± 0.173**	**4.622 ± 0.189**	**4.488 ± 0.128**
	housing	SRF	5.432 ± 0.729	3.845 ± 0.379	3.321 ± 0.274
		DIGMM	4.647 ± 0.411	3.898 ± 0.598	3.688 ± 0.192
		GORF(OURS)	**3.739 ± 0.360**	**3.474 ± 0.330**	**3.164 ± 0.252**

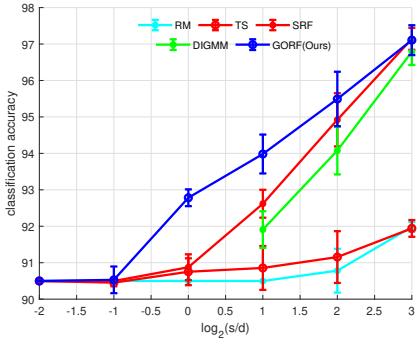
19004035



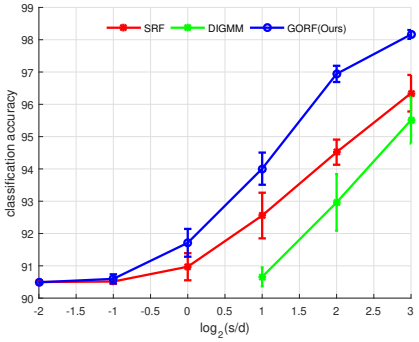
(a) *letter* 数据集，Epanechnikov 核函数 ($a=3, m=1$)



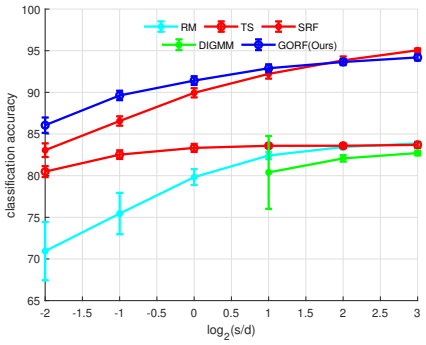
(b) *letter* 数据集，Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=10$)



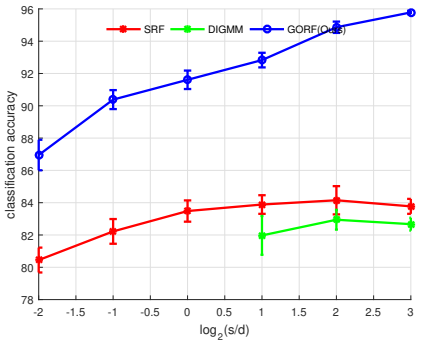
(c) *ijcnn*1 数据集，Epanechnikov 核函数 ($a=3, m=1$)



(d) *ijcnn*1 数据集，Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=10$)



(e) *usps* 数据集，Epanechnikov 核函数 ($a=3, m=1$)



(f) *usps* 数据集，Delta-Gaussian 核函数 ($a_1=1, a_2=-1, \sigma_1=1, \sigma_2=10$)

图 2-6　多种不定核或非平移不变核的随机逼近方法在不同核函数与数据集下 SVM 问题分类准确率的对比实验

Fig. 2-6　Contrast experiments on the SVM classification accuracy of a variety of random approximation methods for various indefinite kernels or non-translation invariant kernels under different kernel functions and datasets

19004035

2.5 本章小结

本章研究了针对不定核和非平移不变核的随机傅里叶特征逼近方法。当所有数据通过归一化约束在单位球面上的时候，非平移不变核函数可以转变为平移不变但非正定的核函数，不定核函数和非平移不变核函数的随机特征逼近构建可以统一成平移不变但非正定核函数的随机特征逼近。对于不定核函数，其频域 $p(\omega)$ 不是非负测度，因此无法通过 Monte-Carlo 采样的方法采集到权值。针对这个问题，本章从符号测度和 Jordan 分解的角度出发，提出了在复数空间中构建不定核的随机傅里叶特征逼近。本章理论证明了提出的逼近方法的无偏性质，并且采用正交化的方法减小逼近的方差，计算采用正交化后减小的方差数值。本章对于不定核的正定分解存在性的前沿问题，从不定核的随机傅里叶特征逼近的角度给出了更加容易验证的充要条件。最后，本章以 Epanechnikov 核函数与 Delta-Gaussian 核函数两个具体的不定核函数，在多个数据集上对比了提出的方法和现有的方法针对不定核函数的逼近性能，验证了提出的方法能够取得更低的逼近误差，在 SVM 和 SVR 问题上具有更好的性能。

19004035

第三章　基于低比特量化随机特征的深层核学习研究

第二章提出的针对不定核函数的低维随机特征逼近突破了随机傅里叶特征正定性和平移不变性的约束，使得能够在更广泛的核函数范围利用低维随机特征逼近对核方法进行加速。然而相关研究果表明核函数需要更强的复杂度来保证更好的泛化性能，对应到随机特征方法而言就是要保证随机特征的数目。相较于直接在单层随机特征逼近基础上增加特征的数目，采用随机特征级联的方法构造的深层核函数具有两个随机特征数目增加的维度，从而使其拥有更强的灵活性和模型表示能力。但是随机特征数目增加会使得存储的采样值矩阵规模增大，限制了随机特征逼近方法在移动端和嵌入式等存储资源有限的场景下的应用。针对这个问题，本章从另一种“低维”的角度，即随机逼近的低比特的表示，来思考在计算存储资源有限的情况下的核函数的随机特征逼近问题。本章设计了针对基于低比特量化的随机特征的深层核函数的构建和训练方法，并在 EEG 数据集进行了实验验证。实验结果证明，低比特量化表示的随机特征能够突破计算存储资源的瓶颈限制，在相同计算存储资源的情况下低比特量化随机特征允许深层核函数的宽度和深度增加保证核方法的泛化能力。

3.1　基于随机特征的深层核构建方法

第二章中的 Bohner 定理说明对于正定核函数，存在非负测度 $p(\omega)$，使得通过 Monte Carlo 抽样得到的采样值矩阵 ω 来构建出随机傅里叶特征，随机傅里叶特征相当于给出了核函数 $k(.,.)$ 对应的近似的特征映射 $\phi(.)$。深层核函数是核函数的多层复合结构，其有多种实现方式。本章采用的是 Xie 等人[57] 提出的基于随机特征的深层核构建方法（DKR: Deep Kernel via Random Fourier Features）。这个方法认为，每层核函数 $k_i(.,.)$ 都有对应的傅里叶变换 $p_i(\omega)$ 和随机傅里叶特征 $\phi_i(.)$，通过随机傅里叶特征 $\phi_i(.)$ 的复合可以构成深层核函数。在 DKR 中，每层的采样值矩阵 $\{\omega_i\}_{i=1}^{l}$ 是可以训练的参数，可以根据 $\{\omega_i\}_{i=1}^{l}$ 构建每层核函数的频谱分布 $\{p_i(\omega)\}_{i=1}^{l}$。通过优化损失函数 $L(w,\phi_l(...\phi_2(\phi_1(x_j;\omega_1))),b,y_j)$，数据驱动学习 $\{\omega_i\}_{i=1}^{l}$，可以

19004035

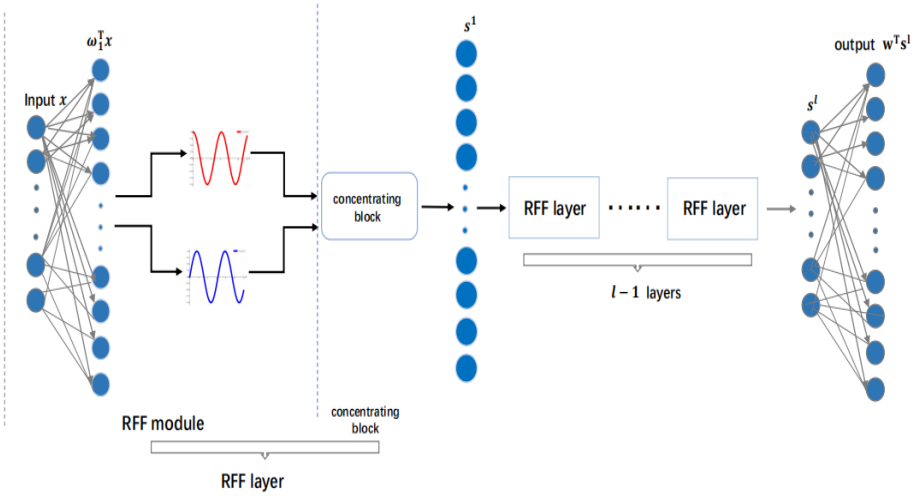


图 3-1　基于随机特征的深层核函数构建方法示意图[57]

Fig. 3-1　The construction method of deep kernel via random features [57]

学习到每层的核函数的频谱分布 $\{p_i(\omega)\}_{i=1}^{l}$ 和核函数 $\{k_i(.)\}_{i=1}^{l}$，即

$$\min_{w,b,\{\omega\}_{i=1}^{l}} \frac{1}{N}\sum_{j=1}^{N} L(w,\phi_l(...\phi_2(\phi_1(x_j;\omega_1))),b,y_j) \tag{3.1}$$

在 SVM 二分类问题中，损失函数 $L(.)$ 是(1.16)的 Hinge 损失函数；同样地，$L(.)$ 也可以是交叉熵损失函数。

图3-1展示的是 DKR 方法的整体框架，DKR 由 l 个 RFF 层（RFF layer）组成，RFF 层含有 RFF 模块（RFF module）和聚合模块（concentrating module）。RFF 模块的主要作用是将上一层核函数的随机傅里叶特征通过采样值矩阵 ω 进行线性变换，产生的结果通过 sin(.) 和 cos(.) 激活函数进行计算，获得随机傅里叶特征的各个分量。聚合模块是将这一层核函数的随机傅里叶特征的分量聚合在一起。经过 l 个的 RFF 层，提取的随机特征最后通过线性分类器进行分类得到样本的预测值。DKR 方法采用梯度下降的方法求取 w，b 和 $\{\omega_i\}_{i=1}^{l}$ 的最优解，设 L_j 表示对于样本 x_j 计算的损失

19004035

函数值，则对 ω_l 的分量 ω_l^m 的导数为：

$$\frac{\partial L_j}{\partial \omega_l^m} = \frac{\partial L_j}{\partial \tilde{y}_j} \cdot \left(\frac{\partial \tilde{y}_j}{\partial s_l^m} \cdot \frac{\partial s_l^m}{\partial \omega_l^m} + \frac{\partial \tilde{y}_j}{\partial s_l^{m+D_l}} \cdot \frac{\partial s_l^{m+D_l}}{\partial \omega_l^m} \right) \tag{3.2}$$

其中 D_l 表示 ω_l 的维度，$\tilde{y}_j = w^\top s_l + b$，

$$\frac{\partial \tilde{y}_j}{\partial s_l^m} = w_m, \quad \frac{\partial \tilde{y}_j}{\partial s_I^{m+D_l}} = w_{m+D_l} \tag{3.3}$$

$$\frac{\partial s_l^m}{\partial \omega_l^m} = -s_{l-1} \sin\left[(\omega_l^m)^\top s_{l-1} \right] \tag{3.4}$$

$$\frac{\partial s_l^{m+D_l}}{\partial \omega_l^m} = s_{l-1} \cos\left[(\omega_l^m)^\top s_{l-1} \right] \tag{3.5}$$

需要注意的是，由于高斯核函数在众多任务中的性能有一定的保证，本方法采用了基于高斯的随机初始化方法对每一层的 ω 进行初始化，即每一层使用了高斯核函数。Xie 等人从 Betti 数的角度对 DKR 方法的模型复杂度进行了分析并且指出在多个小规模和大规模数据集上，DKR 方法的性能远优于单层核学习方法，甚至优于多层感知机（MLP：Multiple Layer Perception）。

3.2 随机特征的低比特量化方法

DKR 方法从随机傅里叶特征级联的角度为深层核函数构造提供了新的方式，假设 DKR 方法中构造的深层核函数中级联随机傅里叶特征的层数为 l，每层随机傅里叶特征的维度为 D_l，数据输入 x 的维度为 D_s。如果 $\{\omega_i\}_{i=1}^l$ 都是以 32 比特进行表示，则消耗的存储容量为：

$$Memory\ Consumption = 32 \times (2D_l^2 \times (l-1) + D_l \times D_x)\ (Bits) \tag{3.6}$$

实际算法部署的过程中经常遇到计算存储资源的限制，如在移动端上往往多个应用程序并行运行，这个时候单个应用程序所需要的内存资源需要进一步压缩。计算存储资源的进一步限制要求对 DKR 的进一步压缩。

19004035

DKR 的压缩有两个思路，一个思路是减小使用的随机特征的数目，即减小单层随机特征的维度或者是减小深层核函数的深度。然而在大规模数据集上，**直接减少所使用的随机特征的数目是否是最佳选择呢**？Avron 等人[86]提出的针对核矩阵的 (Δ_1,Δ_2) 谱逼近理论指出，如果在大规模数据集上保证核方法的泛化性能，需要保证随机特征的采样值矩阵的秩足够高。

定义 3.1. ((Δ_1,Δ_2) **谱逼近**[86]) 对于 $\Delta_1,\Delta_2 \geq 0$，一个对称矩阵 A 是另外一个对称矩阵 B 的 (Δ_1,Δ_2) 谱逼近的充要条件是 $(1-\Delta_1)B \preceq A \preceq (1+\Delta_2)B$。

定理 3.1. [86] 假设 f_K 是核岭回归问题中最优的回归函数，$\tilde{K}+\lambda I$ 是 $K+\lambda I$ 的 (Δ_1,Δ_2) 谱逼近，其中 $\Delta_1 \in [0,1)$ 并且 $\Delta_2 \geq 0$。定义 m 是 $\tilde{K}$ 的秩，并且 $f_{\tilde{K}}$ 是采用核矩阵 $\tilde{K}$ 对应的回归函数，$\lambda \geq 0$ 是正则化项并且标签噪声为 $\sigma^2 < \infty$，则 $f_{\tilde{K}}$ 的期望风险为：

$$\mathcal{R}(f_{\tilde{K}}) \leq \frac{1}{1-\Delta_1}\widehat{\mathcal{R}}(f_K) + \frac{\Delta_2}{1+\Delta_2}\frac{m}{N}\sigma^2 \tag{3.7}$$

当数据规模 N 很大的时候，$f_{\tilde{K}}$ 的期望风险的上界 $\mathcal{R}(f_{\tilde{K}})$ 由 $\frac{1}{1-\Delta_1}\widehat{\mathcal{R}}(f_K)$ 决定。Δ_1 对于 $f_{\tilde{K}}$ 的泛化性能有重要作用，Δ_1 越小，$f_{\tilde{K}}$ 的期望误差越小。Avron 等人通过实验现象指出 RFF 近似的核矩阵 $\tilde{K}$ 的秩越大，Δ_1 的取值越小。而在 RFF 中，逼近的核矩阵 $\tilde{K}$ 的秩等于所有样本 RFF 组合成的逼近矩阵 Z 的秩，而 Z 的秩 $rank(Z) \leq \min(n,s)$，即 Z 的秩的上界由随机特征采样数目 s 决定。定理3.1说明保证通过基于随机特征的核方法的泛化性能需要保证随机特征的采样数目 s，因此直接减少所使用的随机特征的数目在计算存储资源有限的应用场景不是合适的选择。

除了随机特征的数目减少，内存消耗计算公式(3.6)提供了另外一个方向，即 $\{\omega_i\}_{i=1}^{l}$ 都是以小于 32 位的低比特进行表示。本章从**随机特征低比特量化**的角度研究计算存储资源限制条件下随机特征的构建问题，本节主要介绍针随机特征的低比特量化方法及对应的深层核训练方法。

3.2.1 量化函数设计

针对随机特征的低比特量化主要指的是对采样值矩阵 $\{\omega_i\}_{i=1}^{l}$ 和随机特征 $\{s_i\}_{i=1}^{l}$ 两个部分进行低比特表示，从而最大程度地压缩级联的随机特

19004035

征的模型体积，降低运算时间。量化函数是低比特量化的重要组成部分，为了减小量化误差，本章采用的量化函数包含两个部分，分别是量化和反量化。

- **量化部分**

  量化部分将浮点数值转换为定点数表示，给定 32 位浮点数的数值 x，则经过量化部分的 32 位浮点数转变为定点数值 $\bar{x}^q$：

$$\bar{x}^q = \text{clip}\left(\text{round}\left(\frac{x}{c}\right) - z, N_{\min}, N_{\max}\right) \tag{3.8}$$

  其中 round(.) 是取整函数，clip(x,$N_{\min}$, $N_{\max}$) 函数指的是 x 限制在 $N_{\min}$ 和 $N_{\max}$ 之间。$N_{\min}$ 和 $N_{\max}$ 是量化的上界和下界，z 是量化的零点，c 是量化的步长。量化步长 c 的计算公式如下：

$$c = \frac{x_{\max} - x_{\min}}{N_{\max} - N_{\min}} \tag{3.9}$$

  若 $N_{\min} = 0$，$N_{\max} = 2^b - 1$，则

$$c = \frac{x_{\max} - x_{\min}}{2^b - 1} \tag{3.10}$$

  量化零点 z 的计算公式为：

$$z = \frac{x_{\min}}{c} \tag{3.11}$$

  其中 $x_{\min}$ 和 $x_{\max}$ 是 x 的最小值和最大值。

- **反量化部分**

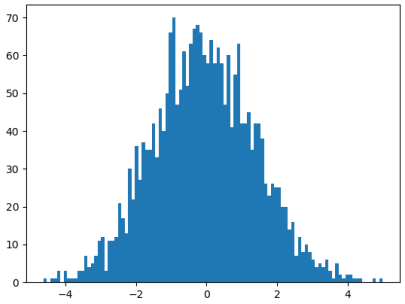
  反量化部分的作用是将定点数 $\bar{x}^q$ 投射回离散的浮点数 x^q，即

$$x^q = c \cdot (\bar{x}^q + z) \tag{3.12}$$

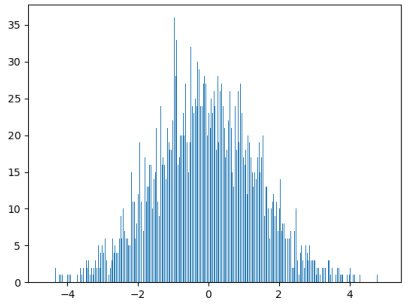
对于逐层采样值矩阵 $\{\omega_i\}_{i=1}^{l}$ 进行 b 比特量化，量化步长 $c_{\omega_i} = \frac{\omega_{i\max} - \omega_{i\min}}{2^b - 1}$，量化零点 $z_{\omega_i} = \frac{\omega_{i\min}}{c_{\omega_i}}$；对于逐层随机特征 $\{s_i\}_{i=1}^{l}$ 进行 b 比特量化，量化步长

19004035

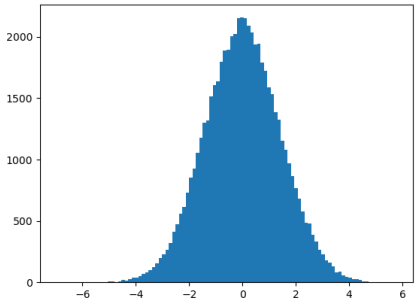
$c_{s_i}=\frac{s_{i\max}-s_{i\min}}{2^b-1}$，量化零点 $z_{s_i}=\frac{s_{i\min}}{c_{s_i}}$。随机傅里叶特征各个分量的值域为 [-1,1]，则 $s_{i\min}=-1$，$s_{i\max}=1$。图3-2展示的是构建的六层 8 比特量化随机特征在第一层和第五层量化前后采样值矩阵 $\boldsymbol{\omega}$ 的分布，通过公式(3.8)和(3.12)，连续分布的采样值矩阵 $\boldsymbol{\omega}$ 量化成离散分布的采样值矩阵 $\boldsymbol{\omega}^q$。



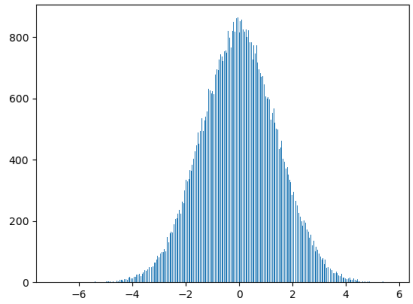
(a) 量化前的第一层采样值矩阵 ω_1



(b) 量化后的第一层采样值矩阵 ω_1^q



(c) 量化前的第五层采样值矩阵 ω_5



(d) 量化后的第五层采样值矩阵 ω_5^q

图 3-2　8 比特量化的 DKR 的每层量化前后采样值矩阵 $\boldsymbol{\omega}$ 的分布

Fig. 3-2　Distribution of sample matrix $\boldsymbol{\omega}$ before and after 8-bit quantization for DKR

3.2.2　基于低比特量化随机特征的深层核训练算法

和基于 32 比特浮点数随机特征的深层核一样，基于低比特量化随机特征的深层核采用了梯度下降的方法对于 w、b 和 $\{\omega_i\}_{i=1}^l$ 进行优化。其中量化函数中的取整函数 round(.) 在量化边界处不可导并且在大多数取值处的导数为 0，基于量化随机特征的深层核函数在训练过程中出现梯度消失或

19004035

爆炸的现象。因此针对量化函数中的求导采用了梯度直传（STE: Straight-through Estimator）的方法，即

$$\frac{\partial x^q}{\partial x} = \mathbb{I}_{|x|\leq 1} \tag{3.13}$$

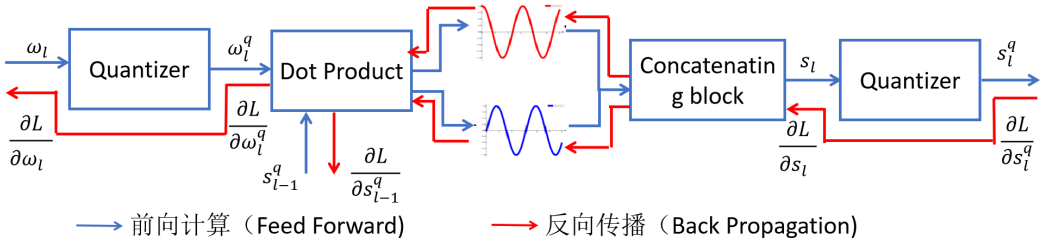


图 3-3　低比特量化的 RFF 层的计算图

Fig. 3-3　Computational graph of low-bit quantized RFF layer

图3-3是低比特量化的 RFF 层的计算图。基于量化随机特征的深层核的训练算法主要分为三个步骤，分别是前向计算、反向梯度传播与采样值矩阵的更新，下面分别对各个部分进行详细地描述。

1) **前向计算**

如图3-3所示，红色箭头代表了低比特量化的 RFF 层前向计算的过程。给定上一层低比特量化后的随机特征 s_{i-1}^q，算法首先将第 i 个 RFF 层的采样值矩阵 ω_i 通过公式(3.8)和(3.12)转化为离散的浮点数 ω_i^q，然后将低比特量化随机特征 s_{i-1}^q 和量化后的采样值矩阵 ω_i^q 进行乘加运算，通过 sin(.) 和 cos(.) 激活函数和聚合模块得到第 i 层的随机傅里叶特征 s_i。最后，随机傅里叶特征 s_i 通过量化函数量化成 s_i^q。重复上述步骤，依次计算逐层的随机傅里叶特征 $\{s_i\}_{i=1}^l$。线性分类器将最后一层得到的随机傅里叶特征 s_l 作为模型的输入，利用分类器的输出和真实的标签求取分类的损失函数 $L(w, \phi_l(...\phi_2(\phi_1(x_j;\omega_1))), b, y_j)$。

2) **反向梯度传播**

图3-3的蓝色箭头是低比特量化的 RFF 层的反向梯度传播的过程。前向计算的过程中涉及到逐层量化前后的采样值矩阵 $\{\omega_i\}_{i=1}^l$ 和 $\{\omega_i^q\}_{i=1}^l$，以及逐层量化前后的随机傅里叶特征输出 $\{s_i\}_{i=1}^l$ 和 $\{s_i^q\}_{i=1}^l$。根据 STE 方法，

19004035

第 j 个样本 x_j 计算的模型输出 $\tilde{y}_j$ 对第 i 层量化前的随机傅里叶特征 s_i 和反向梯度传播中对第 i 层量化后的随机傅里叶特征 s_i^q 相等，即

$$\frac{\partial \tilde{y}_j}{\partial s_i} = \frac{\partial \tilde{y}_j}{\partial s_i^q}\frac{\partial s_i^q}{\partial s_i} = \frac{\partial L^j}{\partial s_i^q}\mathbb{I}_{|s_i|\leq 1} \tag{3.14}$$

根据 3.1 节中的公式 (3.2)-(3.5) 以及公式(3.14)，可以求取损失函数 $L(w,\phi_l(...\phi_2(\phi_1(x_j;\omega_1))),b,y_j)$ 对量化后的采样值矩阵 $\{\omega_i^q\}_{i=1}^l$ 的导数。

3) **采样值矩阵的更新**

反向梯度传播过程求取了损失函数 $L(w,\phi_l(...\phi_2(\phi_1(x_j;\omega_1))),b,y_j)$ 对量化后的逐层采样值矩阵 $\{\omega_i^q\}_{i=1}^l$ 的导数，根据梯度下降方法，可以计算第 t 个迭代周期更新的逐层采样值矩阵 $\{\omega_i\}_{i=1}^l$，即

$$\omega_i(t) = \omega_i(t-1) - \eta\frac{1}{N}\sum_{j=1}^{N}\frac{\partial L_j}{\partial \omega_i^q} \tag{3.15}$$

注意的是公式(3.15)梯度更新过程中采用的是损失函数对量化后的逐层采样值矩阵 $\{\omega_i^q\}_{i=1}^l$ 的梯度，更新的是量化前的逐层采样值矩阵 $\{\omega_i\}_{i=1}^l$。这种采样值矩阵更新方式不仅最小化损失函数 L_j，同时最小化量化 $\{\omega_i\}_{i=1}^l$ 和 $\{s_i\}_{i=1}^l$ 带来的误差。低比特量化的 DKR 训练算法可以总结成为伪代码算法2。

在低比特量化的 DKR 模型训练完成，进行模型部署的时候，量化前的采样值矩阵 $\{\omega_i\}_{i=1}^l$ 被舍弃，保留量化后的采样值矩阵 $\{\omega_i^q\}_{i=1}^l$，此时 DKR 消耗的存储空间将极大地缩小。

3.3 时间和空间复杂度分析

设采样值矩阵 $\{\omega_i\}_{i=1}^l$ 和随机特征 $\{s_i\}_{i=1}^l$ 的量化比特数目为 b 比特，则基于 b 比特量化随机特征的深层核函数模型的空间复杂度为 $\mathcal{O}(b\times(2D_l^2\times(l-1)+D_l\times D_x))$。关于时间复杂度，第 i 层离散浮点数表示的采样值矩阵 ω_i^q 与第 $i-1$ 层离散浮点数表示的随机特征 s_i^q 求取点积，即有：

$$\omega_i^q\cdot s_i^q = c_{\omega_i}c_{s_i}(\bar{\omega}_i^q\cdot\bar{s}_i^q + z_{\omega_i}\bar{s}_i^q + z_{s_i}\bar{\omega}_i^q + z_{\omega_i}z_{s_i}) \tag{3.16}$$

19004035

算法 2 基于低比特量化 RFF 的深层核训练算法

输入: 一个批次的训练数据 $\{x_i, y_i\}_{i=1}^{N}$，每层采样值矩阵 $\{\omega_i\}_{i=1}^{l}$ 的初始分布 $\{p_i(\omega)\}_{i=1}^{l}$，每层的 RFF 的数目 D_l，训练迭代次数 t_{max}

输出: 分类超平面的参数 w 和 b，每层量化后的采样值矩阵 $\{\omega_i^q\}_{i=1}^{l}$

根据初始分布 $\{p_i(\omega)\}_{i=1}^{l}$ 初始化逐层的采样值矩阵 $\{\omega_i\}_{i=1}^{l}$。
for k=1 to t_{max} do
  for j=1 to l do
    对于 $j>1$，根据公式(3.8)和(3.12)将第 j 个 RFF 层的浮点数输入 s_{j-1} 量化成离散浮点数输入 s_{j-1}^q，对于 $j=1$，RFF 层的输入是 x_i
    根据公式(3.8)和(3.12)将第 j 个 RFF 层的采样值矩阵 ω_j 量化成离散浮点数类型的采样值矩阵 ω_j^q
    根据 ω_j^q 和 s_{j-1}^q，计算第 j 层的 RFF 为 s_j
  end for
  采用第 l 层的 RFF 输入线性分类器得到预测值 $\{\tilde{y}_i\}_{i=1}^{N}$，计算损失函数 $L(w, s_l, b, y_j)$
  按照公式 (3.2)-(3.5) 和 (3.13)，求取损失函数 $L(w, s_l, b, y_j)$ 对各层量化后的采样值矩阵 $\{\omega_i^q\}_{i=1}^{l}$ 的梯度，和对线性分类器参数 w 和 b 的梯度。
  根据公式(3.15)更新 $\{\omega_i\}_{i=1}^{l}$，同时根据梯度下降方法更新线性分类器参数 w 和 b
end for

其中 $\bar{\omega}_i^q$ 和 $\bar{s}_i^q$ 是定点量化的采样值矩阵和随机特征，取值为 $[0, 2^b-1]$ 范围中的所有整数值。$\bar{\omega}_i^q$ 表示成 M 位定点整数序列集合，即 $\bar{\omega}_i^q = \sum_{m=0}^{M-1} c_m(\bar{\omega}_i^q)$；$\bar{s}_i^q$ 表示成 K 位定点整数序列集合，即 $\bar{s}_i^q = \sum_{k=0}^{K-1} c_k(\bar{s}_i^q)$，其中 $\left(c_m(\bar{\omega}_i^q)\right)_{m=0}^{M-1}$ 和 $\left(c_k(\bar{s}_i^q)\right)_{k=0}^{K-1}$ 都是位向量。$\bar{\omega}_i^q$ 和 $\bar{s}_i^q$ 的点积可以采用位运算计算，即：

$$\begin{aligned} \bar{\omega}_i^q \cdot \bar{s}_i^q = \sum_{m=0}^{M-1}\sum_{k=0}^{K-1} 2^{m+k} \text{ bitcount } \left[\text{and}\left(c_m(\bar{\omega}_i^q), c_k(\bar{s}_i^q)\right)\right] \\ c_m(\bar{\omega}_i^q)_j, c_k(\bar{s}_i^q)_j \in \{0,1\} \ \forall j,m,k \end{aligned} \tag{3.17}$$

根据公式(3.17)，本章得到 DKR 方法的时间复杂度为 $\mathcal{O}(MK)$，和采样值矩阵及随机特征量化的比特数目有关。当采样值矩阵和随机特征量化比特均为 b 比特时，相较于浮点数表示，时间复杂度下降为原来的 $\frac{b^2}{32^2}$，存储复杂度下降为原来的 $\frac{b}{32}$。

19004035

3.4 实验验证

本小节对基于低比特量化随机特征的 DKR 方法进行实验验证，从不同存储限制下的 DKR 方法的泛化性能角度，对比了基于浮点数随机特征和低比特轻量化随机特征的 DKR 方法。本节所有实验的硬件环境是装有 Ubuntu18.04 系统的塔式服务器，其 CPU 是 24 核 3.5GHz 的 i9-10920X，其 GPU 是显存 11GB 的 GeForce RTX 2080-Ti。实验的软件环境为 Pytorch1.9.0。

3.4.1 不同存储限制下低比特量化 DKR 在 EEG 数据集上的结果

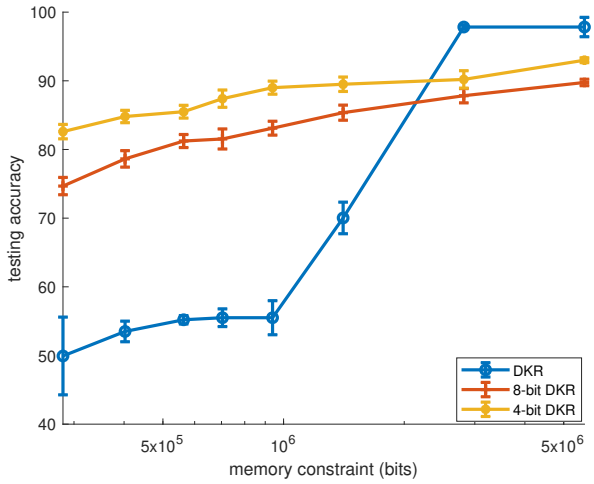
在计算存储资源瓶颈限制下保证 RFF 逼近矩阵的规模是低比特量化随机特征相较于浮点数的随机特征的主要区别，而本小节主要从实验的角度说明这个区别是否导致 DKR 的泛化性能产生差别。实验中采用了 EEG 数据集进行验证，该数据集记录了一段时间受试者脑电信号（EEG：Electroencephalogram) 与眼睛活动情况。数据输入是 14 维的脑电信号数据，输出是受试者的眼睛睁开和闭上的 0/1 变量。数据的规模是 14980，其中随机划分 50% 的数据为训练集，剩下 50% 的数据集构成测试集。实验针对 DKR 进行了两种处理，一种是采用的是 32 位浮点数的 DKR，另一种是采用了低比特量化的 DKR，其中本小节实验中量化比特的数目为 4 和 8。

实验中采用了 7 层 RFF 构成的深层核函数。假定每层 RFF 的维度相同，为了简化运算，本章实验不考虑第一层 RFF 带来的存储负担。根据公式(3.6)，给定存储资源的限制（Memory Constraint），DKR 采用 b 比特情况下对应的每层 RFF 的维度为：

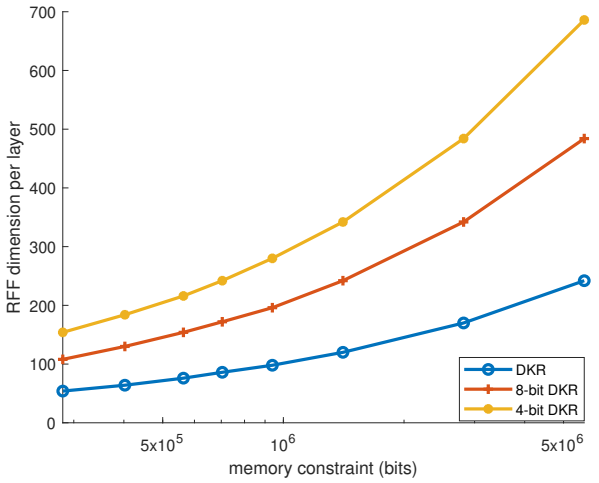
$$2D_l = 2\sqrt{\frac{Memory\ Constraint}{b \times 2 \times (l-1)}} \tag{3.18}$$

实验中研究了 10^5 至 5×10^6 比特存储限制范围内基于 32 位浮点数、4 比特和 8 比特量化 RFF 的深层核函数在 EEG 测试集上分类准确率的情况，如图3-4所示。根据公式(3.18)，我们可以计算得到基于 32 位浮点数、4 比特和 8 比特量化 RFF 的深层核函数在每个存储资源限制点处对应的每层 RFF 维度。如图3-4(b) 所示，4 比特和 8 比特量化允许每层拥有更高维

19004035



(a) EEG 测试数据集上的分类准确率



(b) 每层允许最大的 RFF 的维度

图 3-4　不同存储限制下基于 32 比特浮点数、8 比特和 4 比特量化的 DKR 方法每层允许最大 RFF 的维度以及对应的测试集准确率对比

Fig. 3-4　The maximum RFF dimension of each layer and the testing accuracy for DKR method with 32-bit floating point, 8-bit and 4-bit quantization under different storage restrictions

19004035

度的 RFF。根据每个存储资源限制点处对应的单层 RFF 维度，本小节构建出基于 32 位浮点数，4 和 8 比特量化 RFF 的深层核函数。对于两种数据精度表示的深层核函数，我们在每个存储资源限制点上进行了 5 次重复实验，计算 5 次重复实验测试集分类准确率的均值和标准差。所有实验采用 Hinge 损失函数和 Adam 优化器进行训练，学习率设置为 0.001，训练迭代周期为 1000。

如图3-4所示，当能够提供的存储资源下降到 1.4×10^6 比特时，采用基于 32 位浮点数的 DKR 方法在 EEG 测试集上的分类准确率出现了显著的下降，而采用基于 4 和 8 比特量化的 DKR 方法的性能没有显著的损失。结合图3-4(b) 中每层采用的 RFF 变化曲线，这个现象出现的原因是基于 32 位浮点数的 DKR 方法每层允许最大的 RFF 维度低于保证 DKR 方法泛化性能所要求的 RFF 维度的阈值；而采用 4 比特和 8 比特量化的 DKR 方法允许每层采用更多的 RFF，在一定存储限制范围内保证了 DKR 方法的泛化性能。

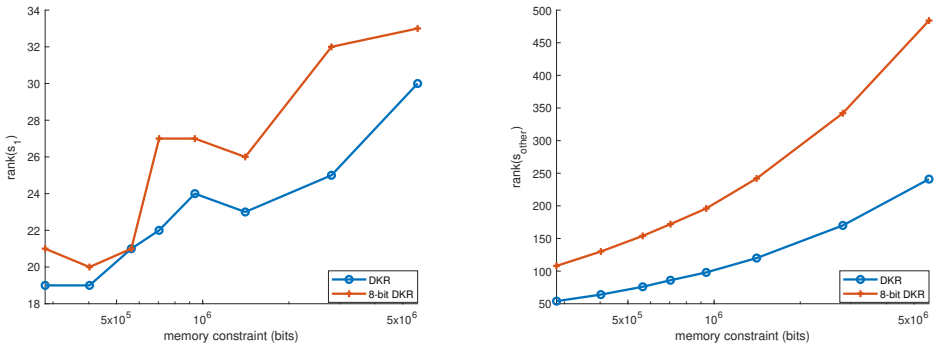
3.4.2 低比特量化 DKR 的进一步讨论

下面将从逐层 RFF 的可视化以及逐层所有样本 RFF 组合成的逼近矩阵 Z 的秩变化这两个角度，以 8 比特量化的 RFF 为例子，进一步分析基于低比特量化的 DKR 构建。

- **不同存储限制下逼近矩阵 Z 的秩的变化**

  Avron 等人提出的 (Δ_1,Δ_2) 谱逼近理论指出基于随机特征的核方法需要通过维持逼近矩阵 Z 的秩足够高来保证核方法的泛化性能。图3-5比较了基于 32 位比特浮点数和 8-比特量化的 DKR 的每层逼近矩阵 Z 的秩。随着存储资源的减小，每层允许的最大 RFF 维度逐渐减小，不同层 RFF 对应的逼近矩阵 Z 的秩都减小。低比特量化不会减小逼近矩阵 Z 的秩，并且由于其允许每层采用更大的 RFF 维度，在相同存储资源限制情况下，8 比特量化 DKR 各层 RFF 对应的逼近矩阵 Z 的秩相比于 32 比特浮点数 DKR 各层 RFF 对应的逼近矩阵 Z 的秩更大。根据 (Δ_1,Δ_2) 谱逼近理论，采用低比特量化 DKR 的核方法拥有更强的泛化性能，这个能够解释图3-4(a) 中基于 8 比特量化的 DKR

19004035



(a) 第一层逼近矩阵 Z 的秩　　(b) 其他层逼近矩阵 Z 的秩

图 3-5　不同存储限制下基于 32 比特浮点数和基于 8 比特量化的 DKR 的每层逼近矩阵 Z 的秩的变化情况

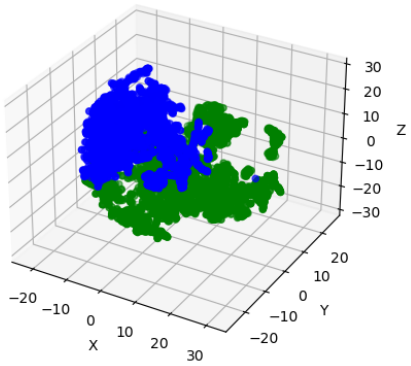
Fig. 3-5　The rank of the approximation matrix Z for DKR method with 32-bit floating point and 8-bit quantization under different storage restrictions

方法在 EEG 测试集上的分类准确率优于基于 32 比特浮点数的 DKR 方法。

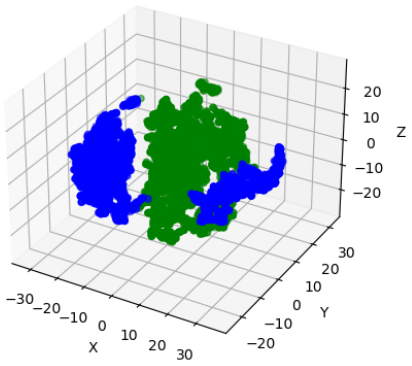
- **逐层 RFF 的可视化**

  RFF 是核函数的无偏逼近，其本质上也是有限维特征映射函数。图3-6从在给定存储资源限制的情况下可视化基于 32 比特浮点数和基于 8 比特量化的 DKR 的第 2、3、4 层 RFF，图中的蓝色和绿色散点代表的是 EEG 测试集的两类样本。可视化方法采用了 t-SNE 降维的方式，t-SNE 降维方法能够在低维分布中保留高维分布的局部特征。实验中给定的存储资源限制为 1.4×10^6 比特，这个条件下基于 32 比特浮点数的 DKR 在 EEG 测试集上的分类准确率为 70.03%，而基于 8 比特量化的 DKR 的分类准确率为 85.37%。可视化的结果显示基于 8 比特量化的 DKR 中不同类别数据的 RFF 特征的可分性更强，而基于 32 比特浮点数的 DKR 中不同类别数据的 RFF 特征具有一定程度的混杂。可视化结果说明低比特量化 RFF 在相同的存储条件下允许更高维度的 RFF，而相比于 RFF 的数据表示精度，RFF 的维度对于 DKR 方法的泛化性能更加重要。

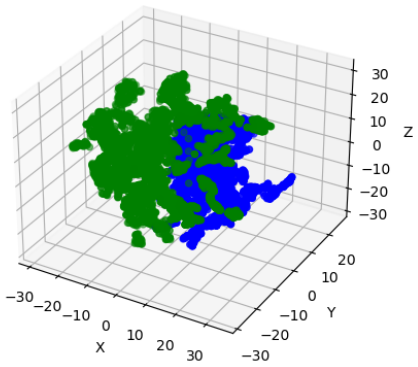
19004035



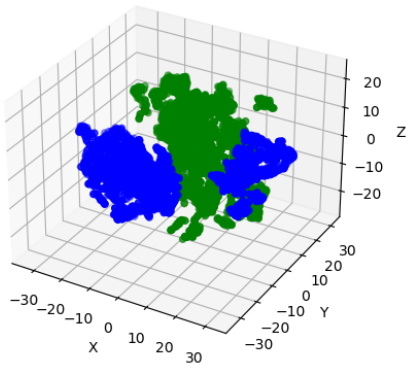
(a) 32 比特 DKR 第 2 层 RFF



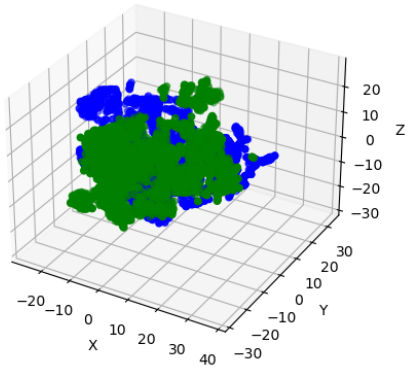
(b) 8 比特 DKR 第 2 层 RFF



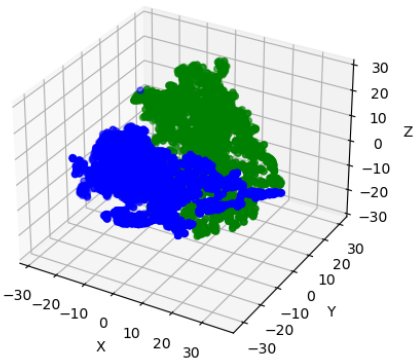
(c) 32 比特 DKR 第 3 层 RFF



(d) 8 比特 DKR 第 3 层 RFF



(e) 32 比特 DKR 第 4 层 RFF



(f) 8 比特 DKR 第 4 层 RFF

图 3-6　存储容量 1.4×10^6 比特下基于 32 比特浮点数和基于 8 比特量化的 DKR 的第 2、3、4 层 RFF 的可视化

Fig. 3-6　Visualization of RFF of layer 2, 3,and 4 for DKR method with 32-bit floating point and 8-bit quantization under the storage capacity of 1.4×10^6 bits

19004035

3.5 本章小结

本章以基于随机特征的深层核函数为研究对象，研究了在计算存储资源约束下的随机傅里叶特征逼近的构建方法。本章根据 Avron 等人提出的理论指出在移动端、嵌入式等计算存储资源有限的情况下，通过减少随机特征的维度来减少存储消耗的方法会导致核方法的泛化性能下降。针对这个问题，本章提出了基于低比特量化随机特征的深层核学习框架，其中包含量化函数的构造及深层核函数的训练方法。在存储资源有限的情况下，低比特量化随机特征允许深层核的宽度和深度增加，来保证基于随机特征的深层核方法的泛化性能。本文理论分析了采用低比特量化随机特征后模型的运算和空间复杂度，相较于 32 位浮点数表示，采用低比特量化后模型的运算和空间复杂度有显著的下降。最后，本章在 EEG 数据集上对比了不同计算存储资源下基于 4 比特和 8 比特量化及基于 32 位浮点数的 DKR 方法，实验结果说明在计算存储资源较少的情况下，低比特量化表示能够较好地保证 DKR 方法的泛化性能。本章分别从逼近矩阵的秩及逐层随机特征可视化两个角度，对低比特量化随机特征能够保证 DKR 方法的泛化性能的原因进行了进一步的讨论。

19004035

19004035

第四章　基于 NTK 低维假设的小样本学习研究

凭借其强大的拟合能力，深度神经网络在计算机视觉、语音语义理解等任务中取得优异的性能结果，然而其训练的过程依赖于大规模数据集的构建。在小规模数据集上，由于优化空间较复杂，深度神经网络的泛化误差增大，导致过拟合的情况。利用先验条件寻找合适的神经网络低维优化空间是提升深度神经网络在小规模数据集上性能的有效方法，而低维优化空间如何构建成为这类方法的核心问题。2018 年，Arthur 等人提出了神经正切核（NTK：Neural Tangent Kernel）[90] 来构建核方法与神经网络训练的动态过程之间的联系。随机特征反映的是核函数能够被低维逼近，Li 等人[107] 根据 NTK 低维逼近的假设，推导出深度神经网络的训练优化过程在低维的子空间中展开。本章利用 NTK 低维逼近假设得到的空间构建小样本学习的优化空间，实验验证在此低维优化空间下，深度神经网络的过拟合现象能够有效的缓解。由于优化子空间无法直接在相同类别对应的大规模数据集上提取，本章提出了通过元学习的方法学习对于不同小样本学习任务迁移性较好的低维优化空间，并在小样本分类任务上进行了实验验证。

4.1　理论背景

作为对于深度神经网络收敛性和泛化性理解的重要工具，NTK 相关研究近年来的数量增长迅速，成为核函数与深度神经网络领域的桥梁。作为本章内容的出发点，本章首先对 NTK 进行介绍，阐明 NTK 如何反映神经网络训练过程，同时从 NTK 低维逼近假设推导出深度神经网络的训练过程是在低维子空间中进行的。

4.1.1　NTK 低维逼近假设

定义深度神经网络的输出 $f(\boldsymbol{\theta},x)\in\mathbb{R}$，其中深度神经网络的参数 $\boldsymbol{\theta}\in\mathbb{R}^m$，深度神经网络的输入 $x\in\mathbb{R}^d$。给定数据集 $\{(x_i,y_i)\}_{i=1}^N\subset\mathbb{R}^d\times\mathbb{R}$，则下面考虑训练数据集上最小化损失函数，假设损失函数采用的是均方误差，即

$$\ell(\boldsymbol{\theta})=\frac{1}{2}\sum_{i=1}^{N}(f(\boldsymbol{\theta},x_i)-y_i)^2 \tag{4.1}$$

19004035

神经网络通过梯度下降的方法进行参数更新，记 t 时刻神经网络的参数为 $\theta(t)$，$t+\Delta t$ 时刻神经网络的参数为 $\theta(t+\Delta t)$，计算 t 时刻损失函数对于网络参数 θ 的导数，即

$$\nabla_\theta \ell(\theta(t)) = \sum_{i=1}^{N} (f(\theta(t),x_i) - y_i) \frac{\partial f((\theta(t),x_i)}{\partial \theta} \tag{4.2}$$

则参数更新的表达式为：

$$\theta(t+\Delta t) = \theta(t) - \eta \nabla_\theta \ell(\theta(t)) \Delta t \tag{4.3}$$

其中 η 是学习速率。当 $\Delta t = 1$ 时，这个对应的是最常用的神经网络参数更新方式。当 Δt 趋近于 0 时，神经网络的参数更新过程可以写成微分方程的形式，即

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta \nabla_\theta \ell(\theta(t)) \tag{4.4}$$

定义 t 时刻对于某个输入样本 x_j 神经网络输出为 $u(t) = f(\theta(t),x_j)$，则神经网络输出 $u(t)$ 的动态变化可以写成：

$$\begin{aligned} \frac{\mathrm{du(t)}}{\mathrm{dt}} &= \frac{\partial \mathrm{u(t)}}{\partial \theta} \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta \frac{\partial \mathrm{u(t)}}{\partial \theta} \nabla_\theta \ell(\theta(t)) \\ &= -\eta \sum_{i=1}^{N} (f(\theta(t),x_i) - y_i) \langle \frac{\partial \mathrm{u(t)}}{\partial \theta}, \frac{\partial f(\theta(t),x_i)}{\partial \theta} \rangle \end{aligned} \tag{4.5}$$

定义 4.1. （**神经正切核**）给定数据集 $\{(x_i,y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ 和深度神经网络 $f(\theta,x)$，其中 x 是深度神经网络的输入，θ 是深度神经网络的参数，随着训练时间 t 会发生变化。则 t 时刻对于样本 x_i 和 x_j 的神经正切核函数定义如下：

$$k(x_i,x_j) = \langle \frac{\partial f(\theta(t),x_i)}{\partial \theta}, \frac{\partial f(\theta(t),x_j)}{\partial \theta} \rangle \tag{4.6}$$

定义矩阵 $K(t)$ 的分量 $K_{ij}(t) = k(x_i,x_j)$，则 K 称为神经正切核矩阵。

公式(4.5)的深度神经网络动态训练过程可以写作矩阵形式，即

$$\frac{\mathrm{du(t)}}{\mathrm{dt}} = -K(t)(u(t) - y) \tag{4.7}$$

19004035

当深度神经网络趋向于无限宽的时候，深度神经网络的参数变化很小，进入“懒惰”模式（lazy training），此时 $K(t)$ 将趋向于恒定的核矩阵 K^*，则深度神经网络动态训练的方程可以写成下面形式，即

$$\frac{\mathrm{du(t)}}{\mathrm{dt}} = -K^*(u(t)-y) \tag{4.8}$$

求解(4.8)的常微分方程，其中常微分方程初值 $u(0)=0$。当 $t\to\infty$ 时，则深度神经网络的输出 $u(t)$ 趋向于

$$f^*(x) = [k(x_1,x),k(x_2,x),...,k(x_N,x)](K^*)^{-1}y \tag{4.9}$$

可以看出在深度神经网络在无限宽的情况下，一个训练收敛的深度神经网络等价于在给定数据集上的核回归预测器，此时可以从核方法的角度对于深度神经网络的特性进行分析。NTK 构建起了核方法与深度神经网络之间的联系。

回顾公式(4.4)，若不考虑损失函数 $\ell(\theta)$ 的具体形式，则可以写做

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta\nabla_\theta\ell(\theta(t)) = -\eta\nabla_\theta(f(\theta(t),\mathcal{X}))^T\nabla_{f(\theta(t),\mathcal{X})}\ell(\theta(t)) \tag{4.10}$$

其中 $\mathcal{X}$ 是数据集所有输入 x 组成的集合，$\nabla_\theta(f(\theta(t),\mathcal{X}))\in\mathbb{R}^{N\times m}$。在无限宽的情况下，深度神经网络对网络参数的梯度保持不变，则

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta\nabla_\theta(f(\theta(0),\mathcal{X}))^T\nabla_{f(\theta(t),\mathcal{X})}\ell(\theta(t)) \tag{4.11}$$

对梯度 $\nabla_\theta(f(\theta(0),\mathcal{X}))$ 采用奇异值分解，可以得到

$$\nabla_\theta(f(\theta(0),\mathcal{X})) = U_0\Sigma_0V_0^T \tag{4.12}$$

其中 $U_0\in\mathbb{R}^{N\times N}$，$V_0\in\mathbb{R}^{m\times m}$ 是实正交矩阵；$\Sigma_0\in\mathbb{R}^{N\times m}$ 是对角矩阵，对角线元素 $\{\lambda_i\}_{i=1,2,...,m}$ 为以递减顺序排列的 $\nabla_\theta(f(\theta(0),\mathcal{X}))$ 的奇异值。定义 NTK 矩阵的特征值为 $\{\lambda_i^{\mathrm{NTK}}\}_{i=1,2,...,m}$，根据 NTK 的定义有 $\lambda_i^{\mathrm{NTK}}=\lambda_i^2$。Fan 等人[108] 对于 NTK 矩阵的特征值分布进行分析，指出 NTK 矩阵中少量的特征值起到主导作用。利用 NTK 矩阵的低秩特性，则可以利用低秩矩

19004035

阵 $\widetilde{\Sigma}_0$ 对于矩阵 Σ_0 进行低秩逼近，即

$$\Sigma_0 \approx \widetilde{U}_0 \widetilde{\Sigma}_0 \widetilde{V}_0^T \tag{4.13}$$

其中 $\widetilde{\Sigma}_0$ 含有 Σ_0 中前 s 大的奇异值，$\widetilde{U}_0 \in \mathbb{R}^{N\times s}$，$\widetilde{V}_0 \in \mathbb{R}^{m\times s}$ 是正交矩阵。则 $\nabla_\theta(f(\theta(0),\mathcal{X}))$ 和 $\frac{\mathrm{d}\theta(t)}{\mathrm{d}t}$ 可以表示成

$$\nabla_\theta f(\mathcal{X},\theta(0)) \approx U_0 \tilde{U}_0 \tilde{\Sigma}_0 \tilde{V}_0^\top V_0^\top \tag{4.14}$$

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta V_0 \tilde{V}_0 [\tilde{\Sigma}_0 \tilde{U}_0^\top U_0^\top \nabla_{f(\theta(t),\mathcal{X})} \ell(\theta(t))] \tag{4.15}$$

其中 $V_0\tilde{V}_0 \in \mathbb{R}^{m\times s}$，$\tilde{\Sigma}_0\tilde{U}_0^\top U_0^\top \in \mathbb{R}^{s\times N}$。$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t}$ 代表的是参数 θ 在训练过程中的变化，公式(4.15)反映训练过程中参数 θ 是在维度为 s 的低维子空间中变化。公式中 $V_0\widetilde{V}_0$ 建立了从 s 维的子空间到 m 维的参数空间的映射，$\tilde{\Sigma}_0\tilde{U}_0^\top U_0^\top \nabla_{f(\theta(t),\mathcal{X})}\ell(\theta(t))$ 指的是在 s 维子空间中投影的梯度。公式(4.15)表明在 s 维的子空间中进行网络训练后最终能够获得在 m 维参数空间一样的性能效果。NTK 低维特性的理论中涉及到两个假设，一个是深度神经网络无限宽，另外一个是深度神经网络的训练进入“懒惰”的状态，这两个条件在实际深度神经网络训练中是理想化的。因此下一小节将从实验现象的角度阐明在真实有限宽深度神经网络中，低维训练的现象也会出现。

4.1.2 深度神经网络低维训练特性

训练深度神经网络到收敛状态，可以获得前 k 次迭代的网络参数值 $\{\theta(1),\theta(2),...,\theta(k)\}$。现在需要考虑的是是否存在和如何找到一个 s 维的子空间，使得深度神经网络的优化过程是在 s 维的子空间中。针对这个问题采用的方法是深度神经网络的参数变化轨迹进行降维，采用的降维方法是主成分分析（PCA：Principal Component Analysis），具体的步骤如下：

1）训练深度神经网络到网络收敛，采集前 k 次迭代的深度神经网络参数值，记为 $\{\theta(1),\theta(2),...,\theta(k)\}$。

2）求取采集的深度神经网络参数的平均值 $\overline{\theta} = \frac{1}{s}\sum_{i=1}^{k}\theta(i)$。

3) 对采集到的深度神经网络参数进行去中心化，即得到 $\Theta = [\theta(1) - \overline{\theta},$

19004035

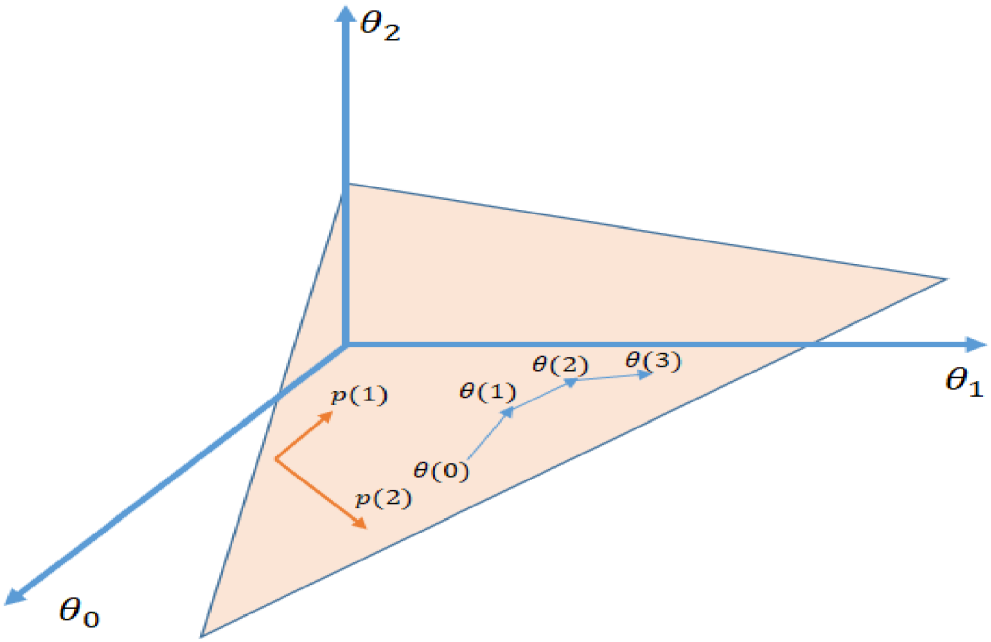


图 4-1　深度神经网络动态训练过程位于低维子空间示意图

Fig. 4-1　The dynamic training process of the deep neural network is located in the low-dimensional subspace

$\theta(2)-\overline{\theta},...,\theta(k)-\overline{\theta}]$

4）对 $\theta^T\theta$ 进行特征值分解，取其中最大的 s 个特征值 $[\sigma_1^2,\sigma_2^2,...,\sigma_s^2]$ 和对应的特征向量 $[\upsilon_1,\upsilon_2,...,\upsilon_s]$。

5）计算降维后的 s 维标准正交基 $P=[p_1,p_2,...p_s]$，其中 $p_i=\frac{1}{\sigma_i}\Theta\upsilon_i$。

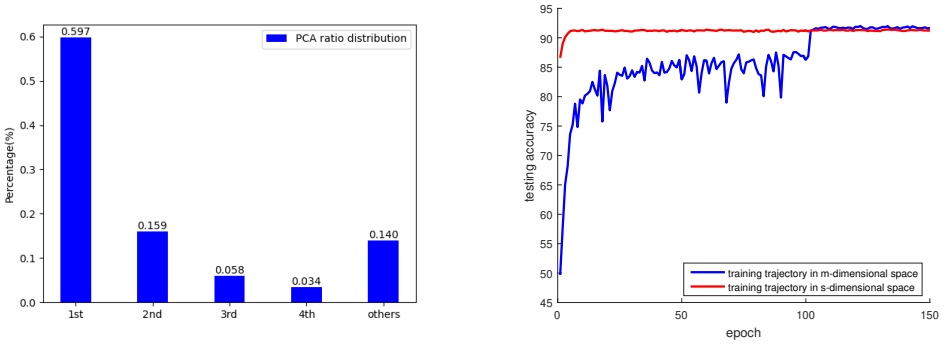
后文将称降维后得到的 s 维标准正交基 P 为投影矩阵。根据 NTK 低维逼近假设，每次迭代过程经过 m 维参数空间到 s 维训练子空间投影以及 s 维训练子空间到 m 维参数空间的变换，深度神经网络最终能够达到和不进行投影过程相同的局部最优点，即

$$f(\theta,x)\approx f(P(P^T(\theta-\overline{\theta}))+\overline{\theta},x) \tag{4.16}$$

19004035

图4-1是深度神经网络的低维训练特性的示意图，即利用(4.16)中的投影矩阵 P 可以构造出深度神经网络在高维参数空间中的低维区域优化的过程。具体在迭代过程中，假设优化器采用的是随机梯度下降算法 (SGD: Stochastic Gradient Descent)，则采用了低维子空间优化后参数更新策略为：

$$\theta(k) = \theta(k-1) - \eta P P^T \nabla_\theta \ell(\theta(k-1)) \tag{4.17}$$



(a) 深度神经网络参数变化轨迹 PCA 降维后各个成分的比例

(b) 深度神经网络在参数空间与低维优化空间测试集准确率曲线

图 4-2　ResNet-20 网络在 CIFAR-10 数据集上采用低维训练策略后的性能

Fig. 4-2　Performance of ResNet-20 network on CIFAR-10 dataset after adopting low-dimensional training strategy.

在本章中，在低维子空间用 SGD 算法进行参数更新的策略称为 P-SGD 算法。图4-2是 ResNet-20 深度神经网络在 CIFAR-10 数据集上采用 SGD 和 P-SGD 优化器的实验结果。首先采用 ResNet-20 网络在 CIFAR-10 数据集上利用 SGD 优化器训练迭代 150 次，测试集上准确率曲线如图4-2(b) 的蓝色曲线所示，最终测试集准确率为 91.64%。接下来对于前 50 个迭代周期的参数变化轨迹降维成 40 维，得到投影矩阵 P。图4-2(a) 是降维后各个主成分所占比例，可以看到最大主成分对于网络参数变化轨迹的方差贡献为 59.7%，说明深度神经网络的参数优化过程是在低维子空间中。重新初始化的 ResNet-20 网络利用 P-SGD 优化器进行重新训练，对应测试集上准确率曲线如图4-2(b) 的红色曲线所示。可以看到，ResNet-20 网络采用 P-SGD 优化算法后不到 10 个迭代周期就收敛，最终测试集准确率为 91.28%，进

19004035

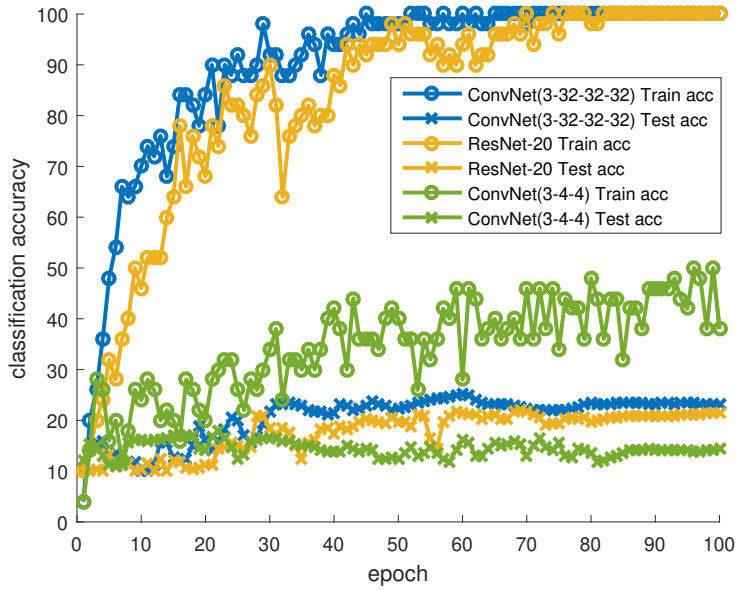


图 4-3　不同深度神经网络在 10-way 5-shot 小规模数据集上训练集和测试集准确率曲线

Fig. 4-3　The training and testing accuracy curves of different deep neural networks on the 10-way 5-shot dataset

一步实验证明了深度神经网络的参数变化轨迹的低维特性。

4.2　小样本学习问题建模

现有深度神经网络能够在实际应用中超越众多统计机器学习方法，CIFAR、ImageNet 等大规模公开数据集起到了重要作用。然而在实际部署中由于数据获取隐私性及人力标注成本等原因，很难采集到大量的数据构成大规模数据集，因此基于小样本的模型学习方法成为学术与工业界中研究的热点。本小节主要对于小样本学习的问题进行数学描述，并且以优化空间的角度对小样本学习的本质进行诠释。

小样本学习问题指的是所拥有的训练数据极少。对于回归问题，信号采样点数目 N 极少，通常采样频率小于信号频率；对于分类问题，其可以描述"N-way K-shot" 的形式，N-way 指的是数据集中有 N 个类别，K-shot 指的是每一类有 K 个样本。其中在小样本学习的研究中 K 通常取值 1 或者 5，样本容量通常少于 50。

19004035

纵然拥有很强的表示能力,深度神经网络在小样本学习问题中容易陷入“过拟合” 的情况，图4-3展示的是三种不同的深度神经网络在 10-way 5-shot 的小样本数据集上的训练集和测试集分类准确率曲线，其中 ConvNet(3-4-4) 和 ConvNet(3-32-32-32) 是两种卷积神经网络，3-4-4 和 3-32-32-32 代表卷积层输入特征的通道数目；ResNet-20 是 20 层残差神经网络，由 Kaiming He 等人[29] 提出。10-way 5-shot 的小样本数据集的构造方法是从 CIFAR-10 数据集的 10 类 50000 张样本中抽取其中每类 5 个样本。可以看到，对于参数量较大的 ConvNet(3-32-32-32) 和 ResNet-20 网络，训练和测试集准确率有比较大的差距，网络陷入严重的 “过拟合” 现象；对于参数量较少的 ConvNet(3-4-4) 网络，训练和测试集准确率之间的差距减小，但由于 ConvNet(3-4-4) 参数量较少，网络表示能力下降，在训练集和测试集上的分类准确率下降。

图4-3出现的现象可以从优化空间的角度进行解释，数据驱动的机器学习问题可以描述成最小化关于数据分布 $p(x,y)$ 和假设（优化）空间 $\mathcal{H}$ 的期望风险，即

$$R(h) = \int \ell(h(x,y)dp(x,y) = \mathbb{E}[\ell(h(x,y))] \tag{4.18}$$

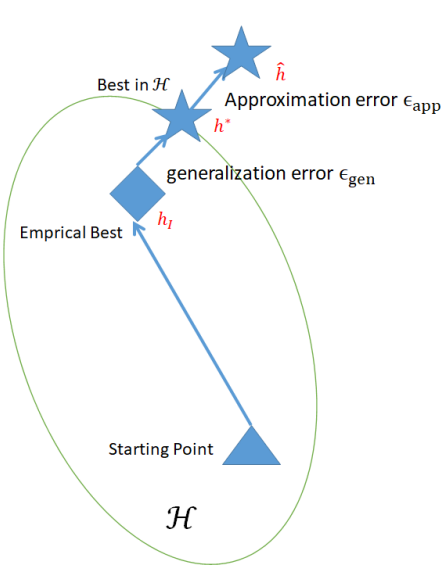
对应深度神经网络 $f(\theta,x)$，$f(.,.)$ 是假设空间 $\mathcal{H}$，而 $f(\theta,x)$ 是假设空间 $\mathcal{H}$ 中的元素。由于数据分布 $p(x,y)$ 未知，实际训练数据集是从数据分布 $p(x,y)$ 进行采样，则深度神经网络的训练过程往往是最小化关于训练数据集和假设空间 $\mathcal{H}$ 的经验风险，即

$$R_I(h) = \frac{1}{N}\sum_{i=1}^{N} \ell(h(x_i,y_i)) \tag{4.19}$$

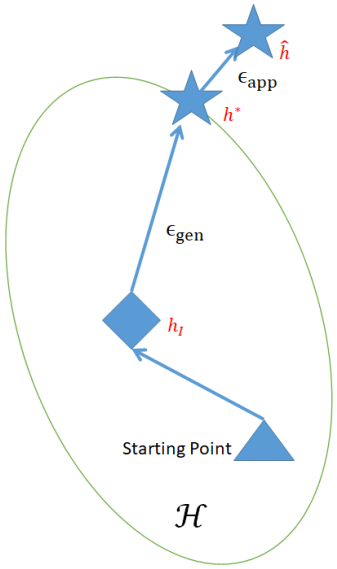
本节中假设 $\hat{h} = argmin_h R(h)$ 是最小化期望误差 $R(h)$ 对应的函数，$h^* = argmin_{h\in\mathcal{H}} R(h)$ 是在假设空间 $\mathcal{H}$ 中最小化期望误差 $R(h)$ 对应的函数，$h_I = argmin_h R_I(h)$ 是在假设空间 $\mathcal{H}$ 中最小化经验误差 $R_I(h)$ 对应的函数。由于无法知道 $\hat{h}$ 的具体形式，因此在机器学习中往往增加假设空间 H 的复杂度来逼近 $\hat{h}$，对应到深度神经网络中就是增大网络的参数量，扩大优化空间。定义关于假设空间 H 的逼近误差 ε_{app} 如下：

$$\varepsilon_{app}(\mathcal{H}) = \mathbb{E}[R(h^*) - R(\hat{h})] \tag{4.20}$$

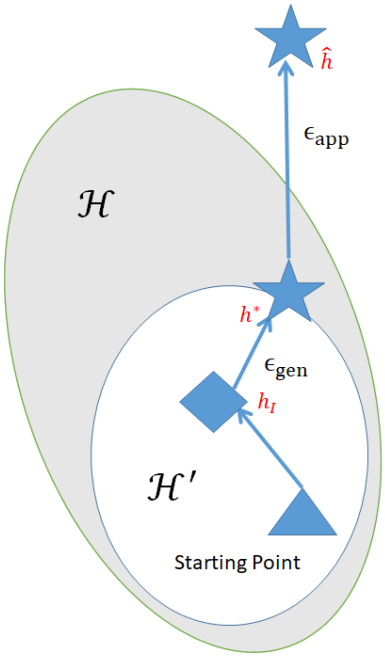
19004035



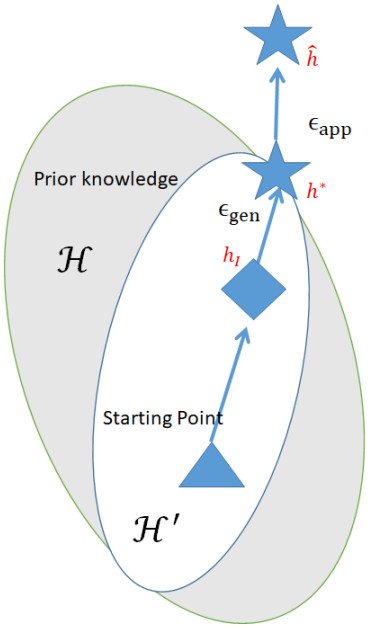
(a) 假设空间 $\mathcal{H}$ 复杂度高，训练数据集规模大



(b) 假设空间 $\mathcal{H}$ 复杂度高，训练数据集规模小



(c) 小样本学习中直接使用参数量少的深度神经网络



(d) 小样本学习中基于先验知识的深度神经网络低维优化空间构建

图 4-4　不同假设（优化）空间复杂度和训练集规模下深度神经网络的优化过程解释

Fig. 4-4　Explanation of the optimization of deep neural networks under different hypothesis (optimization) space complexity and training dataset scale

19004035

假设空间 $\mathcal{H}$ 下经验误差和期望误差的差值构成泛化误差 ε_{gen}，泛化误差不仅与假设空间 $\mathcal{H}$ 的复杂度相关，也与训练数据集相关，即

$$\varepsilon_{gen}(\mathcal{H},\mathcal{X}) = \mathbb{E}[R(h_I) - R(h^*)] \tag{4.21}$$

则关于 h_I 和 $\hat{h}$ 的期望误差的差值可以分解为逼近误差 ε_{app} 和泛化误差 ε_{gen}，即

$$\mathbb{E}[R(h_I) - R(\hat{h})] = \varepsilon_{app}(\mathcal{H}) + \varepsilon_{gen}(\mathcal{H},\mathcal{X}) \tag{4.22}$$

图4-4是不同假设空间复杂度和数据集规模下深度神经网络的优化过程。a) 图展示的是当训练数据集规模很大的时候，由于能够充分地在数据分布 $p(x,y)$ 中进行抽样，因此能够得到 h_I 使得泛化误差 ε_{gen} 小。同时假设空间 $\mathcal{H}$ 复杂度高保证了逼近误差 ε_{app} 小。b) 图对应的小样本学习的情况，当训练数据集规模很小的时候，训练数据集提供的信息有限，较复杂的假设空间使得深度神经网络陷入“过拟合”的情况，此时泛化误差 ε_{gen} 显著增大。一个自然的想法是采用参数量较少的深度神经网络，然而如 c) 图所示，直接裁剪假设空间 $\mathcal{H}$ 至 $\mathcal{H}'$ 会导致模型表示能力减弱，使得逼近误差 ε_{app} 增大，因此直接采用参数量较少的深度神经网络对于小样本学习问题并不是最佳选择。如何根据先验知识构建如 d) 图一样合适的低维优化空间是小样本学习问题的重要研究内容。

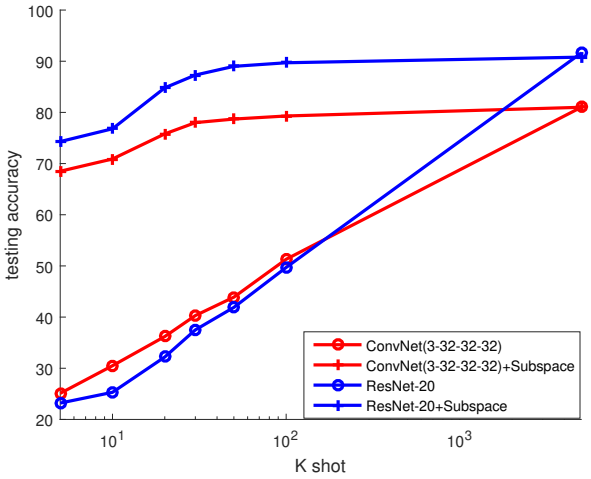
4.3 基于 NTK 低维假设的小样本学习策略

4.3.1 研究动机

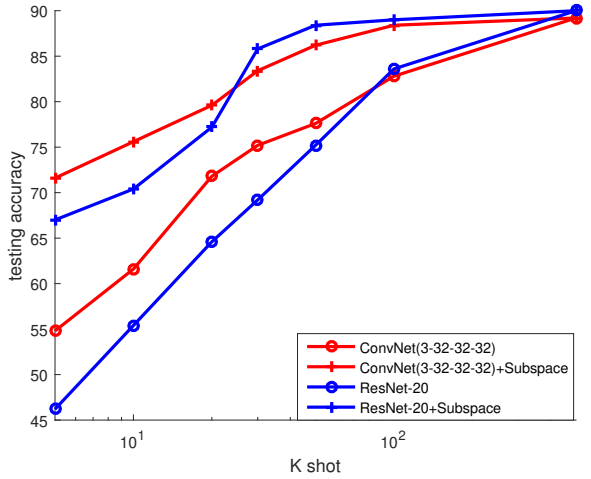
NTK 低维假设推导得到深度神经网络优化并不是在高维的参数空间，而是在由参数的线性组合构成的低维的优化子空间中。NTK 低维假设得到的低维优化子空间能够使得深度神经网络在大规模数据集上快速收敛于相同的性能点，说明提取出来的优化子空间为“最优”的优化子空间。回顾小样本学习，小样本学习需要低维的优化空间，低维优化空间的构造需要先验知识。那么一个很自然的问题产生，NTK **低维假设得到的优化子空间能否作为小样本学习中的低维优化空间呢？**

为了解答这个问题，本小节在不同数据集规模下对比两种深度神经网

19004035



(a) CIFAR-10 数据集



(b) CIFAR-100-small 数据集

图 4-5　不同数据集规模下深度神经网络在高维参数空间和基于 NTK 低维假设的低维子空间训练后测试集准确率

Fig. 4-5　The testing accuracy of the deep neural network optimized in the high-dimensional parameter space and the low-dimensional subspace based on the NTK low-dimensional hypothesis under different dataset scales

19004035

络采用 SGD 和 P-SGD 算法后测试集分类准确率。对比实验使用的两种深度神经网络是 ConvNet(3-32-32-32) 和 ResNet-20，在 CIFAR-10 数据集和 CIFAR-100-small 两种数据集中进行验证，其中 CIFAR-100-small 数据集是从 CIFAR-100 的 100 类中选取 5 类构造而成。数据集的规模由在原始数据集的每类中抽取的样本数目决定。

图4-5是对比结果，其中横坐标 K shot 指的是从数据集的每类中抽取的样本数目，红色曲线代表深度神经网络利用 SGD 算法在原始的高维参数空间训练，蓝色曲线代表深度神经网络利用 P-SGD 算法在基于 NTK 低维假设的低维子空间训练。实验现象表明，当采用 CIFAR-10 和 CIFAR-100-small 所有样本进行训练时，深度神经网络在高维参数空间和基于 NTK 假设的低维子空间训练后测试集准确率接近，这个和 4.1 章节中的深度神经网络低维训练特性这个结论一致的。随着 K-shot 的减小，也就是数据集的规模逐渐减小，可以看到红色曲线和蓝色曲线在测试集准确率上的差距逐渐增大。当数据规模逐步变小的时候，由于训练样本数目不足以支持深度神经网络在复杂的假设空间中优化至图4-4中的 h^* 点，深度神经网络在原始的高维参数空间训练泛化误差增大，“过拟合”情况更加严重；而当深度神经网络在基于 NTK 低维假设的优化子空间中训练的时候，由于训练样本数目和深度神经网络的优化空间的复杂度匹配，并且优化子空间是在 NTK 低维假设的强先验条件下提取的，深度神经网络在小样本学习上出现的“过拟合”情况会得到显著缓解。这种现象说明了 NTK 低维假设得到的优化子空间可以作为小样本学习的低维优化空间，利用这个子空间可以缓解深度神经网络在小规模数据集上的“过拟合”情况。

然而基于 NTK 低维假设的优化空间需要在大规模数据集上提取，即投影矩阵 P 是通过深度神经网络在大规模数据集上训练过程参数变化轨迹 PCA 降维获得，这个强先验条件在小样本学习问题中是理想化的。因此后面两小节在基于 NTK 低维假设的优化空间的基础上，重点解决**小样本学习情况下投影矩阵 P 的学习问题**。

4.3.2 元学习方法

由于小样本学习中和小规模数据集相同类别的大规模数据集无法获得，无法通过直接对参数变化轨迹降维获取降维矩阵 P。针对这个问题，本章

19004035

提出采用元学习的方法在多个同等规模但类别不同的数据集上训练得到对于不同数据集迁移性能较好的降维矩阵 P。本小节重点对元学习方法进行介绍。

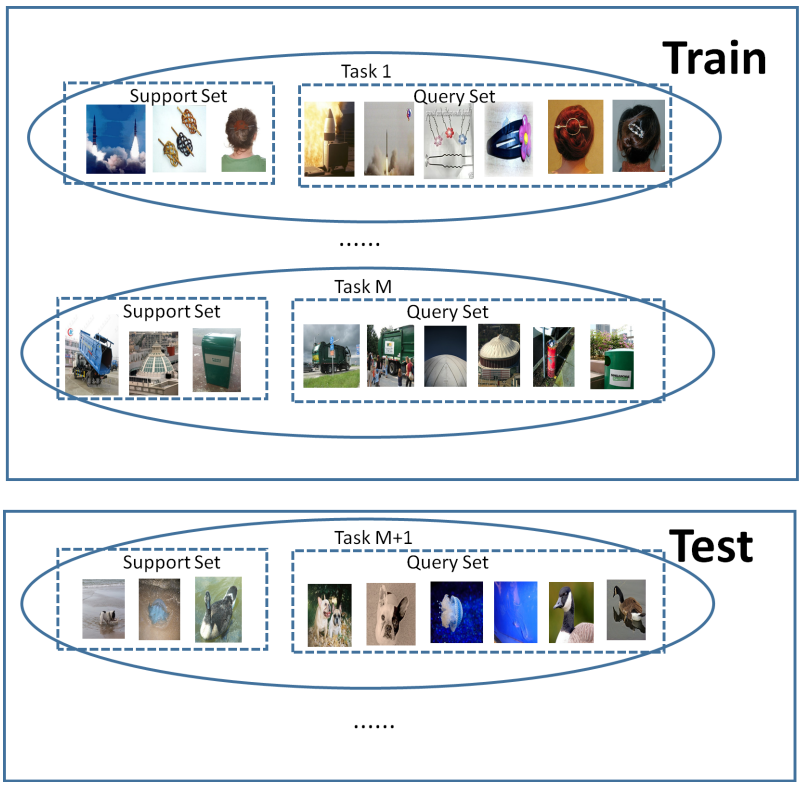


图 4-6　元学习方法对应的小样本学习问题的形式

Fig. 4-6　The few-shot setting in the meta learning method

图4-6是元学习方法对应的小样本学习问题的形式。每个小规模数据集 γ 称为任务（task），在元学习中假设 γ 是从分布 $p(\gamma)$ 中产生。每一个任务 γ 中用于优化基学习器（Base learner）的部分 γ^{tr} 称为支持集合（Support set），用于优化元学习器（Meta learner）的部分 γ^{te} 称为查询集合（Query set）。元学习中所有用于训练的任务构成的集合记为 Γ_{sr}，用于测试的任务构成的集合记为 Γ_{tg}，其中值得注意的是 Γ_{sr} 和 Γ_{tg} 含有的样本类别不同。

元学习方法可以看作是学习一个在多个任务上泛化性能都较好的模型，并且使得模型在新的任务中能够快速适应。和预训练方法（Pre-training）类似，元学习方法需要在不同的数据集中迁移“知识”；和预训练方法的差别

19004035

在于元学习方法是在大量的任务上优化”知识“。元学习方法中称”知识“为元学习器，本小节中记为 ω。元学习器 ω 一般指的是深度神经网络优化过程中的超参数，Fin 等人[109] 提出的 MAML 方法将深度神经网络的初始化参数作为元学习器，Franceschi 等人[110] 和 Li 等人[111] 将正则化项系数及学习率作为元学习器。

元学习方法的模型训练过程是双层（Bi-level）的优化问题，可以数学表示成以下形式：

$$\begin{aligned} \omega^* &= \arg\min_{\omega} \sum_{i=1}^{M} \ell^{meta}(\theta^{*(i)}, \omega, \Gamma_{sr}^{te(i)}) \\ s.t. \quad \theta^{*(i)} &= \arg\min_{\theta} \ell^{task}(\theta, \omega, \Gamma_{sr}^{tr(i)}) \end{aligned} \tag{4.23}$$

其中 ℓ^{meta} 和 ℓ^{task} 是外层和内层优化的目标函数，小样本回归问题通常采用均方误差损失函数，而小样本分类问题采用的是交叉熵损失函数。外层优化元学习器 ω 使得在不同任务上根据元学习器产生的深度神经网络参数 $\theta^{*(i)}$ 都能够取得更好的泛化性能，而内层优化就是在给定元学习器 ω 和任务的情况下，训练深度神经网络参数 θ 使得其在该任务的支持集合上的损失函数值较低。关于元学习研究中外层优化和内层优化采用的最优化算法，MAML 算法和 Reptile 算法中采用的是梯度下降算法。如果优化过程中遇到不可微的部分，在外层优化和内层优化中可以采用强化学习和进化算法。

4.3.3 低维优化子空间的学习算法

投影矩阵 P 是深度神经网络优化过程的超参数，可以作为元学习器。通过元学习过程中的双层优化可以得到在多个新的任务上迁移性较强的低维优化子空间。图4-7是通过元学习方法学习低维优化子空间的示意图，深度神经网络参数采用了随机初始化，投影矩阵 P 采用随机正交矩阵进行初始化。对于 Γ_{sr} 中的每一个任务 γ,深度神经网络 θ 采用公式(4.17)在投影矩阵 P 对应的低维子空间中进行训练，其中内层梯度下降优化的学习率 η 记为 η_{task}。经过 t 次训练迭代，可以得到在 Γ_{sr} 中所有任务上训练后的深度神经网络参数 $\{\theta^i(t)\}_{i=1}^{M}$。根据 $\{\theta^i(t)\}_{i=1}^{M}$ 计算在 Γ_{sr} 中所有任务的

19004035

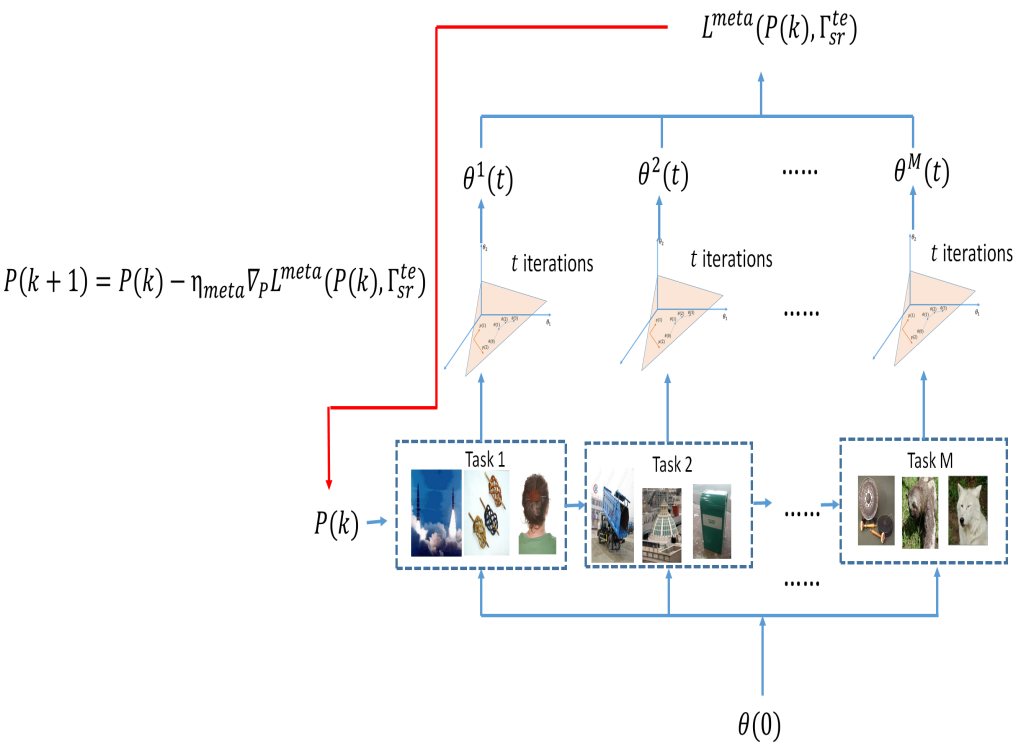


图 4-7　通过元学习方法学习低维优化子空间示意图

Fig. 4-7　The learning process for the low-dimensional optimization subspace through meta learning method

交叉熵损失函数之和 $L^{meta}(P,\Gamma_{sr}^{te})$：

$$L^{meta}(P,\Gamma_{sr}^{te})=\frac{1}{M}\sum_{i=1}^{M}\ell^{meta}(\theta^i(t),P,\Gamma_{sr}^{te^i}) \tag{4.24}$$

根据元学习方法的外部优化过程，接下来需要优化投影矩阵 P 从而最小化 $L^{meta}(P,\Gamma_{sr}^{te})$。由于 $L^{meta}(P,\Gamma_{sr}^{te})$ 中的 $\theta^i(t)$ 是在投影矩阵 P 的作用下通过 P-SGD 优化器训练得到的，$L^{meta}(P,\Gamma_{sr}^{te})$ 对投影矩阵 P 的梯度存在，针对投影矩阵 P 的优化可以采用梯度下降的方法。第 $k+1$ 次迭代对应的投影矩阵 $P(k)$ 为：

$$P(k+1)=P(k)-\eta_{meta}\nabla_P L^{meta}(P,\Gamma_{sr}^{te}) \tag{4.25}$$

19004035

算法 3 基于元学习方法的低维优化子空间的学习方法

输入: 用于模型训练的任务构成的集合 Γ_{sr}，用于模型测试的任务构成的集合 Γ_{tg}，外层优化的学习率 η_{meta} 和优化迭代次数 T_{meta}，内层优化的学习率 η_{task} 和优化迭代次数 T_{task}。

输出: 迁移性较好的低维优化子空间的投影矩阵 P

随机初始化深度神经网络参数 θ 并且采用随机正交矩阵初始化投影矩阵 P，初始化值分别为 $\theta(0)$ 和 $P(0)$
从分布 $p(\gamma)$ 中采集任务 γ 构成 Γ_{sr} 和 Γ_{tg}，从训练任务集合 Γ_{sr} 中采集数据构成 Γ_{sr}^{tr} 和 Γ_{sr}^{te}。
for k=1 to T_{meta} do
    for $t = 1$ to T_{task} do
        对 Γ_{sr}^{tr} 中的所有任务计算深度神经网络的输出 $f(\theta(t-1),\Gamma_{sr}^{tr(i)})$，计算损失函数 $\ell(\theta(t-1))$。
        $\theta(t) = \theta(t-1) - \eta_{task}P(k-1)P(k-1)^T\nabla_\theta\ell'(\theta(t-1))$
    end for
    对 Γ_{sr}^{te} 中的所有任务计算计算深度神经网络的输出 $f(\theta(t-1),\Gamma_{sr}^{te(i)})$，根据公式(4.24)计算损失函数 $L^{meta}(P,\Gamma_{sr}^{te})$
    $P(k) = P(k-1) - \eta_{meta}\nabla_P L^{meta}(P,\Gamma_{sr}^{te})$
end for

模型测试阶段采用的是 Γ_{tg} 上所有任务。此时给定在元学习训练完成的投影矩阵 P，在 Γ_{sr} 的所有任务上通过公式(4.17)训练深度神经网络参数 θ。算法3 将关于基于元学习方法的低维优化子空间的学习方法总结成伪代码。

4.4 实验验证

4.4.1 实验设置

本章在小样本分类任务上验证低维优化子空间对缓解深度神经网络在小样本数据集上的过拟合情况的作用。实验采用了 CIFAR-FS 和 MiniImageNet 两个小样本数据集，下面介绍这两个数据集：

- CIFAR-FS 数据集

  CIFAR-FS 数据集来源于 CIFAR-100 数据集，其中 100 个类别随机划分成 64、16 和 20 个类别用于小样本学习算法的训练、验证和测试。每个类别含有 600 张 32×32 的 RGB 图像。

19004035

- MiniImageNet 数据集

  MiniImageNet 数据集是标准的小样本图像分类数据集，其包含从 ImageNet2012 竞赛数据集 ILSVRC-2012 中随机选取的 100 个类别的图像。100 个类别被随机划分成 64、16 和 20 个类别用于小样本学习算法的训练、验证和测试。每个类别含有 600 张 86×86 的 RGB 图像。

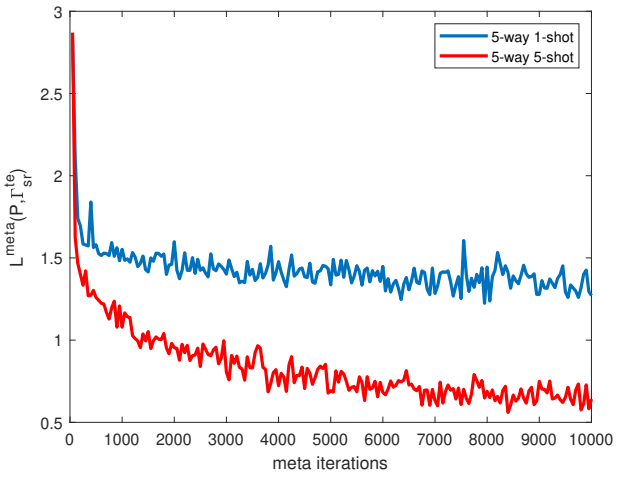
实验验证中采用的深度神经网络结构是四层卷积神经网络 ConvNet(3-32-32-32)，其中 3-32-32-32 是卷积层输入特征的通道数目。4.2 节中提到小样本分类问题采用的是 "N-way K-shot" 的形式，即每一个任务都是 "N-way K-shot"。其中实验中 N 的取值为 $\{5,10\}$，K 的取值为 $\{1,5\}$，N 和 K 组合得到每个任务的训练集样本数目的取值有 $\{5,10,25,50\}$。每个测试集由从每类数据中抽取的 15 个样本构成，当 N 的取值为 5 的时候测试集的规模为 75，当 N 的取值为 10 时测试集的规模为 150。

基于元学习方法得到子空间的过程需要利用到多个任务来对子空间进行优化，元学习有外层和内层两个优化过程。内层利用 P-SGD 优化器对深度神经网络的参数进行优化，学习率 η_{task} 设置为 0.01，每一个任务优化迭代次数 T_{task} 设置为 5；外层利用 Adam 优化器对投影矩阵 P 进行优化，需要的任务数目 M 设置为 16，优化迭代次数 T_{meta} 设置为 20000，学习率 η_{meta} 设置为 0.001。

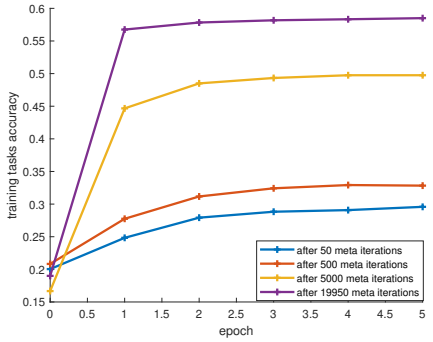
4.4.2 算法的收敛性分析

图4-8展示的是基于 NTK 低维假设与元学习方法的低维子空间方法（meta subspace 方法）两层优化过程，(a) 图描述的是外层优化过程中损失函数 $L^{meta}(P,\Gamma_{sr}^{te})$ 的变化，(b) 图和 (c) 图描述的是 5-way 1-shot 和 5-way 5-shot 两种条件下训练任务的查询集合分类准确率的变化。随着内层优化的进行，(b) 和 (c) 图显示在给定的初始化的低维空间中，深度神经网络参数根据梯度的方向不断优化，在训练任务的查询集合上的分类准确率不断上升。但由于初始化的低维子空间欠佳，分类准确率上升的速度较缓，此时需要对低维子空间进行优化。外层优化针对的是低维子空间的投影矩阵 P，(a) 图显示随着外层优化过程的进行，损失函数 $L^{meta}(P,\Gamma_{sr}^{te})$ 不断下降，低维子空间不断优化，到 5000 个迭代周期收敛到迁移性较好的子空间。深度

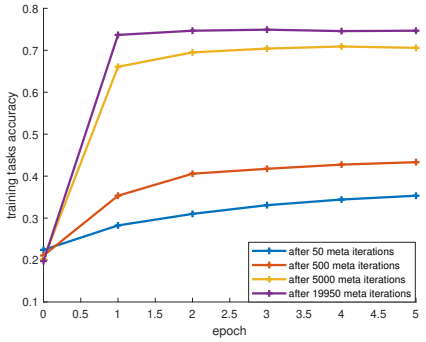
19004035



(a) 外层优化过程中损失函数 $L^{meta}(P,\Gamma_{sr}^{te})$ 的变化



(b) 5-way 1-shot 情况下内层优化中分类准确率的变化

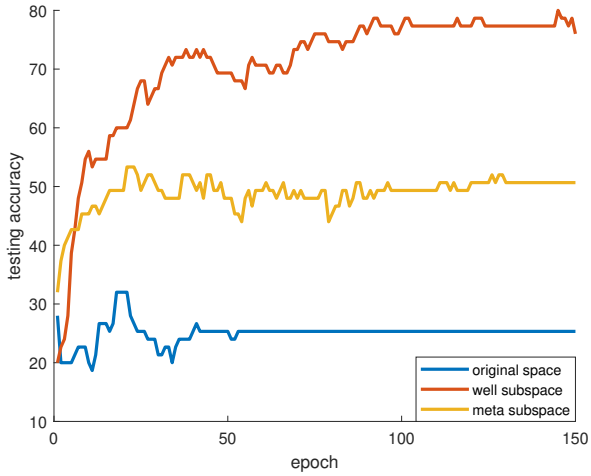


(c) 5-way 5-shot 情况下内层优化中分类准确率的变化

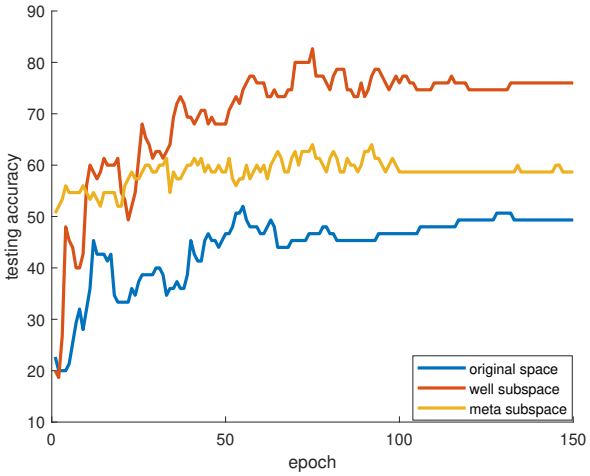
图 4-8　Meta Subspace 方法在 CIFAR-FS 数据集上的双层优化过程

Fig. 4-8　The bi-level optimization process for the meta subspace method on the CIFAR-FS dataset

19004035



(a) 5-way 1-shot



(b) 5-way 5-shot

图 4-9　不同优化空间下 ConvNet 在 CIFAR-FS 测试任务的查询集合上准确率变化曲线

Fig. 4-9　Accuracy curve of ConvNet on the query set of CIFAR-FS testing task under different optimization spaces

19004035

神经网络在不断优化的子空间下收敛速度逐渐加快，并且能够达到更佳的局部最优点，使得其在训练任务的查询集合上的分类准确率升高。

低维子空间优化的优势在于需要学习调整的参数较少，模型搜索空间小，在数据信息有限的情况下可以减缓模型”过拟合“的程度。图4-9展示的是 5=way 1-shot 和 5-way 5-shot 条件下 ConvNet 在 CIFAR-FS 测试任务的查询集合和不同优化空间下分类准确率变化曲线。实验中采对比的优化空间有网络参数原空间 (original space)、基于 NTK 低维假设的理想子空间 (well subspace) 和 meta subspace。选择的 CIFAR-FS 测试任务是 CIFAR-FS 测试任务集合的前 5 个类别，在 original space、well subspace 以及 meta subspace 中采用的学习率 η 均为 0.01。

无论是在 5-way 1-shot 以及 5-way 5-shot 的条件下，ConvNet 在不同优化空间下进行训练后在测试任务的支持集合上分类准确率均可以达到 100%。然而不同优化空间下训练后 ConvNet 在查询集合上的分类准确率差别显著。在 original space 上，ConvNet 训练陷入欠佳的局部最优点，出现了严重的“过拟合”的情况。而在 well subspace 中，ConvNet 可以训练到更好的局部最优点，此时 ConvNet 的“过拟合”情况大幅度缓解。出现这个现象的原因是 well subspace 是根据 ConvNet 在相同类别的大规模数据上参数的训练轨迹提取出来的，虽然减少数据集的规模给 ConvNet 的优化过程提供了更少的信息，但是 well subspace 约束 ConvNet 在理想的低维空间中优化，使得 ConvNet 能够收敛到更加接近全局最优点的位置。然而 well subspace 是理想化的情况，实际应用中获取不到和小样本相同类别的大规模数据集。在 meta subspace 中训练后，ConvNet 在查询集合上达到的准确率显著高于在 original space 中训练的结果，和 well subspace 中训练的结果进一步接近。相较于在 well subspace 中进行训练，ConvNet 在 meta subspace 中训练的时候能够更快地收敛。以上实验结果说明，meta subspace 能够帮助深度神经网络缓解在小样本学习中出现的“过拟合”情况，并且方法具有更强的物理实现性能。

4.4.3 低维优化算法在不同小样本分类数据集上的结果

本小节给出了基于不同优化空间的深度神经网络优化算法在 CIFAR-FS 以及 MiniImageNet 两个小样本分类数据集上的结果。如表4-1所示，pre-

19004035

表 4-1 CIFAR-FS 和 MiniImageNet 数据集上基于不同优化空间的学习算法的分类准确率比较

Table 4-1 Comparison of classification accuracy of learning algorithms based on different optimization spaces on CIFAR-FS and MiniImageNet datasets

数据集	优化空间	5 – way 1 – shot	5 – way 5 – shot	10 – way 1 – shot	10 – way 5 – shot
CIFAR-FS	Original Space	37.95 ± 7.79	53.40 ± 7.29	25.75 ± 3.67	39.52 ± 6.76
	Pretrained Subspace	40.66 ± 6.00	56.25 ± 4.90	21.12 ± 3.17	26.96 ± 4.09
	Well Subspace	64.00 ± 8.94	70.79 ± 8.36	50.39 ± 8.59	64.86 ± 4.17
	Meta Subspace	47.05 ± 8.32	64.55 ± 9.21	36.66 ± 6.30	45.92 ± 5.55
MiniImageNet	Original Space	34.12 ± 8.79	45.55 ± 3.15	20.68 ± 4.46	31.15 ± 5.80
	Pretrained Subspace	36.31 ± 6.22	48.25 ± 5.38	18.31 ± 2.43	24.02 ± 3.18
	Well Subspace	58.24 ± 7.83	64.73 ± 9.56	43.26 ± 6.91	54.52 ± 5.37
	Meta Subspace	44.07 ± 7.65	56.84 ± 8.73	30.06 ± 5.42	37.92 ± 4.75

表 4-2 MiniImageNet 数据集上不同小样本学习算法的分类准确率比较

Table 4-2 Comparison of classification accuracy of different few-shot learning algorithms on the MiniImageNet dataset

小样本学习算法	5 – way 1 – shot	5 – way 5 – shot
ConvNet	34.12 ± 8.79	45.55 ± 3.15
Nearest Neighbor[112]	41.08 ± 6.84	51.04 ± 7.65
Matching Nets[37]	43.56 ± 7.21	55.31 ± 8.31
ConvNet+Meta Subspace	**44.07 ± 7.65**	**56.84 ± 8.73**

trined subspace 指的是基于一个训练任务的参数变化轨迹来提取低维优化空间。实验结果记录的是在 10 个随机选取的测试任务的查询集合上分类准确率的均值和标准差。实验结果显示，在 original space 和 pretrained space 中优化，测试任务的查询集合上分类准确率较低。而在 meta subspace 中优化，测试任务的查询集合上分类准确率显著提升，进一步接近在 well subspace 优化的结果。实验结果说明在 meta subspace 中进行深度神经网络参数的优化可以缓解深度神经网络的”过拟合“的情况。

本小节还对比了基于 meta subspace 的小样本学习方法与部分现有的小样本学习算法在 MiniImageNet 数据集上的分类准确率。其中，Nearest Neighbor 利用了 K-最近邻的思路判断小样本数据集中查询集合的样本和支

19004035

持集合的样本之间的距离,来确定查询集合的样本的标签。Matching Nets[37]利用了神经网络来实现查询集合的样本与支持集合所有样本匹配的过程,通过数据驱动训练神经网络的参数从而学习 Nearest Neighbor 的度量函数。表4-2展示了四种算法在 5-way 1-shot 和 5-way 5-shot 情况下的结果，基于 Meta Subspace 的小样本学习方法在测试任务的查询集合上的准确率高于 Nearest Neighbor 和 Matching Nets 算法。

4.5 本章小结

本章从神经正切核低维假设和深度神经网络低维训练特性的角度思考深度神经网络在小样本学习问题上出现的“过拟合”现象的解决方法。由于优化的参数空间较大并且小规模数据集提供的信息有限，深度神经网络在小样本问题上容易”过拟合“。因此，合适的低维优化空间的构造是小样本学习研究的重要问题。本文从 NTK 低维假设角度出发，得到深度神经网络的优化过程是在低维空间中进行的。随着数据规模减小，深度神经网络在基于 NTK 低维假设得到的优化空间进行训练后泛化性能显著好于在原参数空间训练后的泛化性能，说明基于 NTK 低维假设得到的优化空间使得深度神经网络在小样本学习问题上出现的”过拟合“现象有效缓解。然而基于 NTK 低维假设得到的优化空间需要深度神经网络参数变化的轨迹,属于理想化的先验条件。因此，本章提出通过元学习的方法对优化子空间进行学习和调整。本章在 CIFAR-FS 和 MiniImageNet 两个小样本分类数据集上进行了实验验证，在基于 NTK 低维假设和元学习方法得到的低维空间进行训练时，深度神经网络在小样本数据集上出现的”过拟合“现象得到显著缓解，提出的小样本学习方法在分类准确率上好于 Nearest Neighbor 和 Matching Nets 方法。

19004035

第五章　总结和展望

5.1　全文总结

核函数的随机特征有效地解决了大规模核学习运算和存储复杂度大的问题，并且可以通过级联的结构构建更加灵活的核方法，成为核学习领域的研究热点。本文主要围绕两个角度展开对核函数的随机特征的研究，一个是解决现有核函数的随机特征框架存在的问题，如：核函数的随机特征要求核函数具有正定性和平移不变性，而常用的线性核、多项式核以及用于神经网络动态特性研究的 NTK 等不同时满足这两个性质；计算存储资源有限时，现有随机特征框架直接减少随机特征的维度会带来核方法泛化性能的下降。另一个是探索随机特征及其反映的核函数能够被低维逼近的性质在其他机器学习算法中的应用，如利用核函数的低维逼近研究深度神经网络的低维结构。本文的具体贡献可以归纳为以下几个方面：

1. 在第二章，针对随机傅里叶特征对于核函数的正定性和平移不变性的限制，本文提出了针对复空间下的不定核随机特征逼近方法。当所有数据约束在单位球面上的时候，非平移不变核函数可以转变为平移不变但非正定的核函数。因此，对于不定核和非平移不变核的随机傅里叶特征逼近可以统一为对不定核的随机傅里叶特征逼近。本文从符号测度和 Jordan 分解的角度出发，提出了在复数空间中构建不定核的随机傅里叶特征逼近方法。本文理论证明了提出的逼近方法的无偏性质，并且采用正交化的方法减小逼近的方差。本文以 Epanechnikov 核函数与 Delta-Gaussian 核函数两个具体的不定核函数，在多个数据集上对比了提出的方法和现有针对不定核和非平移不变核函数的逼近性能，验证了提出的方法能够取得更低的逼近误差，在 SVM 和 SVR 等学习问题上具有更好的性能。

2. 在第三章，针对计算存储限制下核函数随机特征的逼近问题，本文提出了低比特量化随机特征的构建方法，以及相应的基于低比特量化的随机特征的深层核学习框架，来缓解减少随机特征维度带来的核方法的泛化性能的损失。相比于 32 位浮点数表示，低比特量化的随机特征允许深层核函数在计算存储资源限制的情况下宽度和深度的增加。本文理论分析低比特量化能够带来的模型压缩和运算加速的程度。最后在不同的计算存储资

19004035

源下本文验证了在存储资源有限的情况下随机特征的低比特轻量化表示方法比 32 位浮点数表示方法更好的学习性能。

3. 第四章是随机特征及其代表的核函数的低维逼近思想在其他机器学习算法中的应用的探索。针对深度神经网络在小样本学习问题上出现的“过拟合”现象，本文基于 NTK 低维假设和深度神经网络的低维训练特性，构造了深度神经网络在小样本学习问题上的低维优化空间。随着数据规模减小，深度神经网络在基于 NTK 低维假设得到的优化空间进行训练后泛化性能显著好于在原参数空间训练后的泛化性能，说明基于 NTK 低维假设得到的优化空间使得深度神经网络在小样本学习问题上出现的”过拟合“现象有效缓解。然而基于 NTK 低维假设得到的优化空间需要深度神经网络参数变化的轨迹，属于理想化的先验条件。因此，本文提出通过元学习的方法对优化子空间进行学习和调整。本文在 CIFAR-FS 和 MiniImageNet 两个小样本分类数据集上进行了实验验证，在基于 NTK 低维假设和元学习方法得到的低维空间进行训练时，深度神经网络在小样本数据集上出现的”过拟合“现象得到显著缓解，并且本文提出的小样本学习方法在分类准确率上好于 Nearest Neighbor 和 Matching Nets 方法。

5.2 未来展望

本文的研究工作存在一定的局限性，部分问题有待下一步更加深入地探究：

1. 本文在第二章提出的核函数的随机傅里叶特征逼近方法突破了正定性和平移不变性的限制，然而目前提出的随机特征逼近方法要求核函数的傅里叶变换具有解析表达式。而许多核函数具有傅里叶变换，但其傅里叶变换只能够数值求解。针对这种类型的核函数，我们需要研究如何设计相应的随机特征的采样方式。

2. 本文在第四章提出了基于低维优化空间的小样本学习方法，低维优化空间通过 NTK 低维假设构造并且利用元学习方法调整。然而，提出方法只是给出了小样本学习的低维优化空间，没有给出在低维优化空间中的合适的优化的初始点及优化的方向。后续工作可以研究如何在低维优化空间中选择合适的初始点和优化的方向，并且结合现有的小样本学习算法，进一步深度神经网络在小样本学习任务上的性能。

19004035

参 考 文 献

[1] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[2] Schölkopf B, Smola A J, Bach F. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. MIT press, 2002.

[3] Drucker H, Burges C J C, Kaufman L, et al. Support vector regression machines[C]. Advances in neural information processing systems, 1997, 9: 155-161.

[4] Schölkopf B, Smola A, Müller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural computation, 1998, 10(5): 1299-1319.

[5] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[C]//Advances in neural information processing systems. 2002: 849-856.

[6] Zhu J, Hastie T. Kernel logistic regression and the import vector machine[J]. Journal of Computational and Graphical Statistics, 2005, 14(1): 185-205.

[7] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning[M]. New York: Springer series in statistics, 2001.

[8] Marinazzo D, Pellicoro M, Stramaglia S. Kernel method for nonlinear Granger causality[J]. Physical review letters, 2008, 100(14): 144103.

[9] Gao S, Tsang I W H, Chia L T. Kernel sparse representation for image classification and face recognition[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 1-14.

19004035

[10] Liu F, Zhou T, Fu K, et al. Kernelized temporal locality learning for real-time visual tracking[J]. Pattern Recognition Letters, 2017, 90: 72-79.

[11] Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification using string kernels[J]. Journal of Machine Learning Research, 2002, 2(Feb): 419-444.

[12] 王敬. 文本分类中 SVM 核函数的探讨 [硕士论文]. 兰州: 兰州大学, 2021.

[13] Rasmussen C E. Gaussian processes in machine learning[C]//Summer school on machine learning. Springer, Berlin, Heidelberg, 2003: 63-71.

[14] Lin C J. Large-scale kernel machines[M]. MIT press, 2007.

[15] Sonnenburg S, Rätsch G, Schäfer C, et al. Large scale multiple kernel learning[J]. The Journal of Machine Learning Research, 2006, 7: 1531-1565.

[16] Zhang Q, Filippi S, Gretton A, et al. Large-scale kernel methods for independence testing[J]. Statistics and Computing, 2018, 28(1): 113-130.

[17] Chen B, Liu H, Bao Z. Optimizing the data-dependent kernel under a unified kernel optimization framework[J]. Pattern recognition, 2008, 41(6): 2107-2119.

[18] Ding M, Tian Z, Xu H. Adaptive kernel principal component analysis[J]. Signal Processing, 2010, 90(5): 1542-1553.

[19] Si S, Hsieh C J, Dhillon I. Memory efficient kernel approximation[C]//International Conference on Machine Learning. PMLR, 2014: 701-709.

[20] Cortes C, Mohri M, Talwalkar A. On the impact of kernel approximation on learning accuracy[C]//Proceedings of the thirteenth international

19004035

conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010: 113-120.

[21] 李俊彬. 核函数逼近方法若干理论与应用研究 [博士论文]. 大连：大连理工大学, 2017.

[22] Rahimi A, Recht B. Random Features for Large-Scale Kernel Machines[C]// Advances in Neural Information Processing Systems. 2007, 1: 1177–1184.

[23] Liu F, Huang X, Chen Y, et al. Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond[J]. arXiv preprint arXiv: 2004.11154, 2020.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[25] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[26] Katharopoulos A, Vyas A, Pappas N, et al. Transformers are rnns: Fast autoregressive transformers with linear attention[C]//International Conference on Machine Learning. PMLR, 2020: 5156-5165.

[27] Choromanski K M, Likhosherstov V, Dohan D, et al. Rethinking Attention with Performers[C]//International Conference on Learning Representations. 2020.

[28] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

19004035

[29] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[30] 刘娟宏, 胡阈, 黄鹤宇. 端到端的深度卷积神经网络语音识别 [J]. 计算机应用与软件, 2020, 37(04):192-196.

[31] Jiang H. Uniform convergence rates for kernel density estimation[C]//International Conference on Machine Learning. PMLR, 2017: 1694-1703.

[32] Shawe-Taylor J, Williams C K I, Cristianini N, et al. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA[J]. IEEE Transactions on Information Theory, 2005, 51(7): 2510-2522.

[33] Belkin M, Hsu D, Xu J. Two models of double descent for weak features[J]. SIAM Journal on Mathematics of Data Science, 2020, 2(4): 1167-1180.

[34] Liu F, Liao Z, Suykens J. Kernel regression in high dimensions: Refined analysis beyond double descent[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2021: 649-657.

[35] 李航. 统计机器学习方法 [M]. 北京：清华大学出版社，2012.

[36] Zhang H, Koniusz P. Zero-shot kernel learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7670-7679.

[37] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[J]. Advances in neural information processing systems, 2016, 29: 3630-3638.

[38] Snell J, Swersky K, Zemel R. Prototypical Networks for Few-shot Learning[J]. Advances in Neural Information Processing Systems, 2017, 30: 4077-4087.

19004035

[39] Dinuzzo F, Schölkopf B. The representer theorem for Hilbert spaces: a necessary and sufficient condition[C]. Advances in Neural Information Processing Systems, 2012, 25: 189-196.

[40] 周志华. 机器学习 [M]. 北京：清华大学出版社，2016.

[41] Hsieh C J, Si S, Dhillon I. A divide-and-conquer solver for kernel support vector machines[C]//International conference on machine learning. PMLR, 2014: 566-574.

[42] Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression[C]//Conference on learning theory. PMLR, 2013: 592-617.

[43] Williams C, Seeger M. Using the Nyström Method to Speed Up Kernel Machines[J]. Advances in Neural Information Processing Systems, 2000: 661-667.

[44] Kumar S, Mohri M, Talwalkar A. Ensemble Nystrom Method[J]. Advances in Neural Information Processing Systems, 2009, 22: 1060-1068.

[45] Oglic D, Gärtner T. Nyström method with kernel k-means++ samples as landmarks[C]//International Conference on Machine Learning. PMLR, 2017: 2652-2660.

[46] Kar P, Karnick H. Random feature maps for dot product kernels[C]//Artificial intelligence and statistics. PMLR, 2012: 583-591.

[47] 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.

[48] Meister M, Sarlos T, Woodruff D. Tight Dimensionality Reduction for Sketching Low Degree Polynomial Kernels[J]. Advances in Neural Information Processing Systems, 2019, 32: 9475-9486.

[49] Shustin P F, Avron H. Gauss-Legendre Features for Gaussian Process Regression[J]. arXiv preprint arXiv:2101.01137, 2021.

19004035

[50] Gönen M, Alpaydın E. Multiple kernel learning algorithms[J]. The Journal of Machine Learning Research, 2011, 12: 2211-2268.

[51] Bucak S S, Jin R, Jain A K. Multiple kernel learning for visual object recognition: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1354-1369.

[52] Cho Y, Saul L. Kernel Methods for Deep Learning[J]. Advances in Neural Information Processing Systems, 2009, 22: 342-350.

[53] Zhuang J, Tsang I W, Hoi S C H. Two-layer multiple kernel learning[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 909-917.

[54] Wilson A G, Hu Z, Salakhutdinov R, et al. Deep kernel learning[C]//Artificial intelligence and statistics. PMLR, 2016: 370-378.

[55] Liu F, Huang X, Gong C, et al. Nonlinear pairwise layer and its training for kernel learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 3659-3666.

[56] Liu F, Huang X, Gong C, et al. Learning Data-adaptive Non-parametric Kernels[J]. Journal of Machine Learning Research, 2020, 21(208): 1-39.

[57] Xie J, Liu F, Wang K, et al. Deep kernel learning via random Fourier features[J]. arXiv preprint arXiv:1910.02660, 2019.

[58] Fang K, Huang X, Liu F, et al. End-to-end Kernel Learning via Generative Random Fourier Features[J]. arXiv preprint arXiv:2009.04614, 2020.

[59] Le Q, Sarlós T, Smola A. Fastfood-approximating kernel expansions in loglinear time[C]//Proceedings of the international conference on machine learning. 2013: 244-252.

19004035

[60] Choromanski K, Sindhwani V. Recycling randomness with structure for sublinear time kernel expansions[C]//International Conference on Machine Learning. PMLR, 2016: 2502-2510.

[61] Feng C, Hu Q, Liao S. Random feature mapping with signed circulant matrix projection[C]//Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015: 3490-3496.

[62] Li P. Linearized GMM kernels and normalized random Fourier features[C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 315-324.

[63] Yu F X X, Suresh A T, Choromanski K M, et al. Orthogonal random features[J]. Advances in neural information processing systems, 2016, 29: 1975-1983.

[64] Bojarski M, Choromanska A, Choromanski K, et al. Structured adaptive and random spinners for fast machine learning computations[C]//Artificial Intelligence and Statistics. PMLR, 2017: 1020-1029.

[65] Choromanski K, Rowland M, Sarlós T, et al. The geometry of random features[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2018: 1-9.

[66] Choromanski K, Rowland M, Chen W, et al. Unifying orthogonal monte carlo methods[C]//International Conference on Machine Learning. PMLR, 2019: 1203-1212.

[67] Yang J, Sindhwani V, Avron H, et al. Quasi-Monte Carlo feature maps for shift-invariant kernels[C]//International Conference on Machine Learning. PMLR, 2014: 485-493.

[68] Lyu Y. Spherical structured feature maps for kernel approximation[C]//International Conference on Machine Learning. PMLR, 2017: 2256-2264.

19004035

[69] Shen W, Yang Z, Wang J. Random features for shift-invariant kernels with moment matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2017: 2520-2526.

[70] Li Z, Ton J F, Oglic D, et al. Towards a unified analysis of random Fourier features[C]//International Conference on Machine Learning. PMLR, 2019: 3905-3914.

[71] Wang Y, Shahrampour S. A general scoring rule for randomized kernel approximation with application to canonical correlation analysis[J]. arXiv preprint arXiv:1910.05384, 2019.

[72] Liu F, Huang X, Chen Y, et al. Random fourier features via fast surrogate leverage weighted sampling[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 4844-4851.

[73] Sinha A, Duchi J C. Learning kernels with random features[J]. Advances in Neural Information Processing Systems, 2016, 29: 1298-1306.

[74] Cortes C, Mohri M, Rostamizadeh A. Two-stage learning kernel algorithms[C]//27th International Conference on Machine Learning, ICML 2010. 2010: 239-246.

[75] Li C L, Chang W C, Mroueh Y, et al. Implicit kernel learning[C]//The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019: 2007-2016.

[76] Yu F X, Kumar S, Rowley H, et al. Compact nonlinear maps and circulant extensions[J]. arXiv preprint arXiv:1503.03893, 2015.

[77] Wilson A, Adams R. Gaussian process kernels for pattern discovery and extrapolation[C]//International conference on machine learning. PMLR, 2013: 1067-1075.

19004035

[78] Smola A J, Ovari Z L, Williamson R C. Regularization with dot-product kernels[J]. Advances in neural information processing systems, 2001: 308-314.

[79] Hinton G. Stochastic neighbor embedding[J]. Advances in neural information processing systems, 2003, 15: 857-864.

[80] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9: 2579-2605.

[81] Kullback S, Leibler R A. On information and sufficiency[J]. The annals of mathematical statistics, 1951, 22(1): 79-86.

[82] 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.

[83] Kafai M, Eshghi K. CROification: accurate kernel classification with the efficiency of sparse linear SVM[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1): 34-48.

[84] Pennington J, Yu F X X, Kumar S. Spherical Random Features for Polynomial Kernels[J]. Advances in Neural Information Processing Systems, 2015, 28: 1846-1854.

[85] Liu F, Huang X, Shi L, et al. A double-variational Bayesian framework in random Fourier features for indefinite kernels[J]. IEEE transactions on neural networks and learning systems, 2019, 31(8): 2965-2979.

[86] Avron H, Kapralov M, Musco C, et al. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees[C]//International Conference on Machine Learning. PMLR, 2017: 253-262.

[87] Tu S, Roelofs R, Venkataraman S, et al. Large scale kernel learning using block coordinate descent[J]. arXiv preprint arXiv:1602.05310, 2016.

19004035

[88] May A, Garakani A B, Lu Z, et al. Kernel approximation methods for speech recognition[J]. The Journal of Machine Learning Research, 2019, 20(1): 2121-2156.

[89] Liao Z, Couillet R, Mahoney M W. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent[J]. Advances in Neural Information Processing Systems, 2020, 33: 13939-13950.

[90] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[J]. arXiv preprint arXiv:1806.07572, 2018.

[91] Arora S, Du S S, Hu W, et al. On Exact Computation with an Infinitely Wide Neural Net[J]. Advances in Neural Information Processing Systems, 2019, 32: 8141-8150.

[92] Han I, Avron H, Shoham N, et al. Random Features for the Neural Tangent Kernel[J]. arXiv preprint arXiv:2104.01351, 2021.

[93] Malach E, Yehudai G, Shalev-Schwartz S, et al. Proving the lottery ticket hypothesis: Pruning is all you need[C]//International Conference on Machine Learning. PMLR, 2020: 6682-6691.

[94] Cohn D L. Measure Theory[M]. Springer Science & Business Media, 2013.

[95] Bochner S. Harmonic Analysis and the Theory of Probability[M]. Courier Corporation, 2005.

[96] Bognár J. Indefinite inner product spaces[M]. Springer Science & Business Media, 2012.

[97] Ong C S, Mary X, Canu S, et al. Learning with non-positive kernels[C]//Proceedings of the twenty-first international conference on Machine learning. 2004: 639-646.

19004035

[98] Graepel T, Herbrich R, Bollmann-Sdorra P, et al. Classification on pairwise proximity data[J]. Advances in neural information processing systems, 1999: 438-444.

[99] Pekalska E, Paclik P, Duin R P W. A generalized kernel approach to dissimilarity-based classification[J]. Journal of machine learning research, 2001, 2(Dec): 175-211.

[100] Huang X, Maier A, Hornegger J, et al. Indefinite kernels in least squares support vector machines and principal component analysis[J]. Applied and Computational Harmonic Analysis, 2017, 43(1): 162-172.

[101] Liu F, Huang X, Gong C, et al. Indefinite kernel logistic regression with concave-inexact-convex procedure[J]. IEEE transactions on neural networks and learning systems, 2018, 30(3): 765-776.

[102] Huang X, Suykens J A K, Wang S, et al. Classification With Truncated ℓ_1 Distance Kernel[J]. IEEE transactions on neural networks and learning systems, 2017, 29(5): 2025-2030.

[103] Alabdulmohsin I, Gao X, Zhang X Z. Support vector machines with indefinite kernels[C]//Asian Conference on Machine Learning. PMLR, 2015: 32-47.

[104] Nikolentzos G, Meladianos P, Vazirgiannis M. Matching node embeddings for graph similarity[C]//Thirty-first AAAI conference on artificial intelligence. 2017: 2429-2435.

[105] 武爱文，冯卫国等. 概率论与数理统计 [M]. 上海：上海交通大学出版社，2016.

[106] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. the Journal of machine Learning research, 2008, 9: 1871-1874.

19004035

[107] Li T, Tan L, Tao Q, et al. Train Deep Neural Networks in 40-D Subspaces[J]. arXiv preprint arXiv:2103.11154, 2021.

[108] Fan Z, Wang Z. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks[J]. Advances in neural information processing systems, 2020, 33: 7710-7721.

[109] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning. PMLR, 2017: 1126-1135.

[110] Franceschi L, Frasconi P, Salzo S, et al. Bilevel programming for hyperparameter optimization and meta-learning[C]//International Conference on Machine Learning. PMLR, 2018: 1568-1577.

[111] Li Y, Yang Y, Zhou W, et al. Feature-critic networks for heterogeneous domain generalization[C]//International Conference on Machine Learning. PMLR, 2019: 3915-3924.

[112] Atkeson C G, Moore A W, Schaal S. Locally Weighted Learning[J]. Artificial Intelligence Review, 1997, 11(1-5):11-73.

19004035

附录

本附录主要包括对文中提出的定理的证明。

A. **定理2.5的证明**

证明.（**必要性**）和 RKKS 相关的不定核函数有正定分解

$$k(x-y)=k_+(x-y)-k_-(x-y)$$

其中 $k_+(x-y)$ 和 $k_-(x-y)$ 是正定核函数。根据 Bohner 定理，其对应的随机傅里叶变换是非负且有限的 Borel 测度 $p_+(\omega)$ 和 $p_-(\omega)$，即有

$$k(z)=k_+(z)-k_-(z)=\int_{\mathbb{R}^d}\exp(\mathrm{i}\omega^T z)p_+(d\omega)-\int_{\mathbb{R}^d}\exp(\mathrm{i}\upsilon^T z)p_-(d\upsilon)$$

由于 $p_+(\omega)$ 和 $p_-(\omega)$ 是非负且有限的 Borel 测度，则

$$||p_+||=\int_0^\infty |p_+(||\omega||)|d||\omega||=\int_0^\infty p_+(||\omega||)d||\omega||<+\infty$$

得到 $p_+(\omega)$ 的总质量是有限的，同理 $p_-(\omega)$ 的总质量也是有限的。则

$$||p||=||p_+||+||p_-||<+\infty$$

$p(\omega)$ 的总质量是有限的。

（**充分性**）令 $\Omega:=\mathbb{R}^d$，$\mathcal{A}$ 是包含 Ω 所有开子集在内的最小 σ-代数。不定核函数 $k(z)$ 的（广义）傅里叶变换为：

$$p(\omega)=\int_\Omega \exp(-\mathrm{i}\omega^{\mathrm{T}}z)\mathrm{k}(z)\mathrm{d}z$$

由于 $p(\omega)$ 的总质量是有限的，即 $||p(\omega)||<\infty$，则 $p(\omega)\in(-\infty,+\infty)$，$p(\omega)$ 可以看做是符号测度。根据 Jordan 定理，符号测度 $p(\omega)$ 可以分解成 $p_+(\omega)$ 和 $p_-(\omega)$。则根据公式(2.20)和(2.23)，此时不定核函数可以分解成两个正定核函数的差。由此可证得不定核函数 k 具有正定分解 $k=k_+-k_-$。

19004035

B. **定理**2.6**的证明**

证明：根据公式(2.20)和公式(2.5)，有

$$\begin{aligned}\mathbb{E}[K_{\mathrm{GRFF}}(z)] &= \mathbb{E}\left[\sum_{i=1}^{s}\frac{1}{s}||p_+||\cos(\omega_i^T z) - \sum_{i=1}^{s}\frac{1}{s}||p_-||\cos(\upsilon_i^T z)\right] \\ &= \sum_{i=1}^{s}\frac{||p_+||}{s}\mathbb{E}(\cos(\omega_i^T z)) - \sum_{i=1}^{s}\frac{||p_-||}{s}\mathbb{E}(\cos(\upsilon_i^T z))\end{aligned}$$

其中 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$ 分别在 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 中进行独立同分布采样。根据 Bohner 定理，有

$$\mathbb{E}(\cos(\omega_i^T z)) = \widetilde{k}_+(z), \quad \mathbb{E}(\cos(\upsilon_i^T z)) = \widetilde{k}_-(z)$$

则公式(5.1)可以写成：

$$\begin{aligned}\mathbb{E}[K_{\mathrm{GRFF}}(z)] &= \sum_{i=1}^{s}\frac{||p_+||}{s}\widetilde{k}_+(z) - \sum_{i=1}^{s}\frac{||p_-||}{s}\widetilde{k}_-(z) \\ &= ||p_+||\widetilde{k}_+(z) - ||p_-||\widetilde{k}_-(z) = k(z)\end{aligned}$$

证明得到 $K_{\mathrm{GRFF}}(z)$ 的无偏性，下面求取逼近方差，由于权值是随机同分布采样，则有：

$$\begin{aligned}\mathbb{V}[K_{\mathrm{GRFF}}(z)] &= \mathbb{V}\left[\sum_{i=1}^{s}\frac{1}{s}||p_+||\cos(\omega_{\mathrm{i}}^{\mathrm{T}} z) - \sum_{\mathrm{i}=1}^{\mathrm{s}}\frac{1}{\mathrm{s}}||\mathrm{p}_-||\cos(\upsilon_{\mathrm{i}}^{\mathrm{T}} z)\right] \\ &= \frac{||p_+||^2}{s^2}\sum_{i=1}^{s}\mathbb{V}[\cos(\omega_i^T z)] + \frac{||p_-||^2}{s^2}\sum_{i=1}^{s}\mathbb{V}[\cos(\upsilon_i^T z)] \\ &= \frac{||p_+||^2}{s}(\mathbb{E}[\cos^2(\omega_1^T z)] - \widetilde{k}_+^2(z)) + \frac{||p_-||^2}{s}(\mathbb{E}[\cos^2(\upsilon_1^T z)] - \widetilde{k}_-^2(z)) \\ &= \frac{||p_+||^2}{s}\left[\frac{1+\widetilde{k}_+(2z)}{2} - \widetilde{k}_+^2(z)\right] + \frac{||p_-||^2}{s}\left[\frac{1+\widetilde{k}_-(2z)}{2} - \widetilde{k}_-^2(z)\right]\end{aligned}$$

C. **定理**2.8**的证明**

证明：设 $a_i = \cos(\omega_{\mathrm{i}}^{\mathrm{T}} z)$，$b_i = \cos(\upsilon_{\mathrm{i}}^{\mathrm{T}} z)$，因为 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$ 正交后不

19004035

再满足独立同分布采样的性质，则

$$\begin{aligned}
&\mathbb{V}(K_{\mathrm{GORF}}(z)) = \mathbb{V}\left[\sum_{i=1}^{s}\frac{1}{s}||p_+||a_i - \sum_{i=1}^{s}\frac{1}{s}||p_-||b_i\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{s}\frac{1}{s}||p_+||a_i - \sum_{i=1}^{s}\frac{1}{s}||p_-||b_i\right)^2\right] - \mathbb{E}^2\left[\sum_{i=1}^{s}\frac{1}{s}||p_+||a_i - \sum_{i=1}^{s}\frac{1}{s}||p_-||b_i\right] \\
&= \frac{||p_+||^2}{s^2}\sum_{i=1}^{s}\left[\mathbb{E}^2(a_i) - \mathbb{E}(a_i^2)\right] + \frac{||p_-||^2}{s^2}\sum_{i=1}^{s}\left[\mathbb{E}^2(b_i) - \mathbb{E}(b_i^2)\right] + \frac{||p_+||^2}{s^2}\sum_{i,j=1,i\neq j}^{s}[\mathbb{E}(a_i a_j) \\
&- \mathbb{E}(a_i)\mathbb{E}(a_j)] + \frac{||p_-||^2}{s^2}\sum_{i,j=1,i\neq j}^{s}[\mathbb{E}(b_i b_j) - \mathbb{E}(b_i^2)] + H(z)
\end{aligned}$$

由于 $\{\omega_i\}_{i=1}^s$ 和 $\{\upsilon_i\}_{i=1}^s$ 是从 $\widetilde{p}_+(\omega)$ 和 $\widetilde{p}_-(\upsilon)$ 中进行采样，则根据定理2.7，有

$$\mathbb{V}(K_{\mathrm{GORF}}(z)) = \mathbb{V}[K_{\mathrm{GRFF}}(z)] + ||p_+||^2 G_{\widetilde{k}_+}(z) + ||p_-||^2 G_{\widetilde{k}_-}(z) + H(z)$$

对于 $H(z)$ 的计算，有

$$\begin{aligned}
H(z) &= \frac{2}{s^2}||p_+||||p_-||[\mathbb{E}[\sum_{i=1}^{s}a_i]\mathbb{E}[\sum_{j=1}^{s}b_j] - \mathbb{E}[\sum_{i,j=1}^{s}a_i b_j]] \\
&= 2||p_+||||p_-||[\mathbb{E}(a_1)\mathbb{E}(b_1) - \mathbb{E}(a_1 b_1)]
\end{aligned}$$

19004035

19004035

攻读学位期间学术论文和科研成果目录

[1] **Luo Q**, Fang K, Yang J, Huang X. Towards Unbiased Random Features with Lower Variance For Stationary Indefinite Kernels[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.（EI，CCF-C **类，交大** A **类会议，已发表**）

[2] Chu T, **Luo Q**, Yang J, Huang X. Mixed-precision quantized neural networks with progressively decreasing bitwidth[J]. Pattern Recognition, 2021, 111: 107647.（SCI **一区，**CCF-B **类期刊，**IF:7.74**，已发表**）

19004035

19004035

攻读学位期间参与的项目

[1] 多层非正定核学习理论、算法及应用研究，国家自然科学基金（No. 61977046)。

[2] 企业项目“电院-美敦力智慧医疗联合实验室”单目内窥镜图像深度重建。

[3] 国防项目“面向人机协同的智能博弈方法研究”。

19004035

19004035

致　　谢

时光荏苒，两年半的硕士生涯转瞬即逝。突如其来的疫情打乱了人们正常的生活节奏，同时悄然塑造了如今新的世界格局，我的硕士生涯在这个时代的大背景下进行的。虽然疫情给整个硕士学习生活造成挑战，但是我自己最终如期完成了在硕士研究生初期给自己设定的“三个目标”，我个人觉得达到了我对于硕士研究生生涯的预期。整个硕士阶段主要起到承上启下的关键作用，“承上”指的是延续自己在上海交大本科四年的时光，继续完成本科四年未尽的事业；“启下”指的是找准自己人生的发展定位，为未来博士生涯打好科学研究的基本功，在更宽广的舞台上发光发热。

说到完成硕士毕业论文要感谢的人，我最先要感谢的是我的导师黄晓霖老师，他严谨认真的治学态度和对于学术研究一贯的热情深深感染了我；在整个研究生过程中，他提供了很多研究上的指导和帮助，使得我对于学术研究拥有了更加深入的理解和认识；在个人未来发展上他也坚定支持我做出的选择。其次我要感谢实验室的杨杰老师、屠恩美老师以及刘方辉师兄，谢谢他们营造了良好的实验室氛围，并且在科研上给我的帮助。

我也要感谢一路与我同行的小伙伴，楚天舒、何凡、方坤、谢佳轩、王凯捷、姚乐宇、吴颖雯、谭雷、李涛、陈思哲、何铭震、李明哲、谢志强等 PAMI 实验室同学，谢谢他们在科研生活上给我的帮助，他们的陪伴让我的硕士生涯不再孤单；感谢宿舍舍友何志宇、吴庶宸同学及好朋友赵晨辰，经常性与他们的人生探讨给了我很多启示，在我迷茫的时候指引了我前行的方向。我还要感谢身边和远方的朋友，他们给了我很多关爱与温暖。最后，我尤其要感谢我的家人，我的母亲、逝去的父亲和我的外婆，他们是最爱我的。我的父亲在我高考的时候去世了，成为了我人生中永远的一道难以抹去的疤痕；我的母亲在我读研期间身体状况逐渐糟糕，我内心有很深的愧疚。但我想，他们的心愿也就是希望我健康快乐地活下去，面对纷繁复杂的世界，我要坚强勇敢地走下去，才能最大可能地实现他们的心愿。

我硕士毕业后即将加入香港中文大学工程学院院长黄定发教授的团队攻读博士学位，这对于我来说是更大的平台和挑战。“今番良晤，豪兴不浅，他日江湖相逢，再当杯酒言欢。咱们就此别过。”交大岁月永远是我心中美好的记忆，希望自己毕业后有机会回到待过将近七年的地方多看看。

19004035

上海交通大学

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。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作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：罗钦

日期：2022年2月17日

上海交通大学

学位论文使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于 ☑公开论文

☐内部论文，☐1年/☐2年/☐3年 解密后适用本授权书。

☐秘密论文，____年（不超过10年）解密后适用本授权书。

☐机密论文，____年（不超过20年）解密后适用本授权书。

（请在以上方框内打“√”）

学位论文作者签名：罗钦　　指导教师签名：

日期：2022年2月17日　　日期：2022年2月17日

上海交通大学硕士学位论文答辩决议书

姓　名	罗钦	学号	119032910073	所在学科	控制工程
指导教师	黄晓霖	答辩日期	2022-02-15	答辩地点	闵和校区电信群楼2-406
论文题目	核函数低维随机特征逼近的广义构建方法及应用研究				

投票表决结果：5/5/5（同意票数/实到委员数/应到委员数） 答辩结论：☑通过 ☐未通过

评语和决议：

统计机器学习中的核学习研究的热点。罗钦同学的硕士论文针对现有随机特征式的限制及空间复杂度大的问题，重点研究随机特征在更广范围的核函数空间下的构建方法，具有重要的理论研究和应用价值。

论文提出了在复数空间下构建不定核函数的随机傅里叶特征，给出逼近方差减小的方法来实现无偏且低方差的逼近，突破了核随机傅里叶特征对于核函数的正定性和平移不变性的约束；提出了在计算存储资源有限的情况下采用低比特量化的随机特征构建方法，保证了低计算存储资源下随机特征数目及相应核方法的泛化能力；针对深度神经网络应用中面临的小样本学习问题，基于神经正切核低维假设和神经网络低维训练特性构造优化空间，并通过元学习的方法对优化空间进行学习和调整，从而有效缓解“过拟合”的发生。

论文逻辑严谨，工作量充实，理论分析和实验设计合理，具有一定的理论创新，表明作者已经掌握本学科的专业知识，具备独立从事科研的能力。答辩中罗钦同学表述清晰，回答问题准确。经答辩委员会认真讨论并无记名投票，一致同意通过罗钦同学工程硕士学位论文答辩，并建议授予其工程硕士学位。

2022 年 2 月 15日

答辩委员会成员签名	职务	姓名	职称	单位	签名
	主席	沈红斌	教授	上海交通大学电子信息与电气工程学院(自动化系)	
	委员	高岳	副教授	上海交通大学	
	委员	赵群飞	教授	上海交通大学电子信息与电气工程学院(自动化系)	
	委员	霍宏	副研究员	上海交通大学电子信息与电气工程学院(自动化系)	
	委员	孙景乐	高级工程师	上海新产业光电技术有限公司	
	秘书	吴沂军	工程师	上海交通大学电子信息与电气工程学院(自动化系)	