Pure Mathematics 理论数学, 2026, 16(6), 142-154
Published Online June 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.166164

Hans汉斯

核梯度下降在回归问题中的泛化误差与谱偏置

马子骁，崔文泉

中国科学技术大学管理学院，安徽 合肥

收稿日期：2026年4月21日；录用日期：2026年5月24日；发布日期：2026年6月30日

摘 要

本文主要研究使用核梯度下降求解最小二乘回归问题，并基于再生核希尔伯特空间的谱分解，给出了核梯度下降的泛化误差在各特征子空间上的模态误差关于优化时间的函数，这种分解方法有助于我们理解核梯度下降对回归函数在各特征空间上的分量的带偏置的学习，即谱偏置，以及噪声对各模态学习不同程度的影响。在使用核方法时，核函数的选择以及核函数超参数的选择尤为重要，我们的结果验证了任务与模型的对齐理论，这将帮助我们选择适合任务的核函数。由于宽神经网络的训练过程等价于使用神经正切核进行核梯度下降，本文的模态误差函数同样适用于此类网络。在推导本文主要结果时，我们用到了协方差算子的谱分解、矩阵指数函数的拉普拉斯逆变换以及样本协方差矩阵的各向异性局部律等方法，并用高斯核以及神经正切核在人工合成数据以及MNIST数据集上验证了本文的结果。

关键词

核梯度下降，谱偏置，神经正切核，宽神经网络

Generalization Error and Spectral Bias of Kernel Gradient Descent in Regression

Zixiao Ma, Wenquan Cui

School of Management, University of Science and Technology of China, Hefei Anhui

Received: April 21, 2026; accepted: May 24, 2026; published: June 30, 2026

Abstract

This paper primarily investigates the use of kernel gradient descent for solving least squares regression problems. Based on the spectral decomposition of reproducing kernel Hilbert spaces, we present the generalization error of kernel gradient descent as a function of optimization time, specifically the mode error on each eigenspace. This decomposition helps us understand the biased learning of the kernel gradient descent on the components of the regression function in different

文章引用: 马子骁, 崔文泉. 核梯度下降在回归问题中的泛化误差与谱偏置[J]. 理论数学, 2026, 16(6): 142-154.
DOI: 10.12677/pm.2026.166164

eigenspaces—referred to as spectral bias—as well as the varying effects of noise on the learning of different modes. The choice of kernel function and its hyperparameters is crucial when applying kernel methods, and our results validate the task-model alignment theory, which aids in selecting appropriate kernel functions for specific tasks. Since the training process of wide neural networks is equivalent to kernel gradient descent using the neural tangent kernel, the mode error function derived in this paper is also applicable to such networks. In deriving the main results, we employ techniques such as the spectral decomposition of the covariance operator, the inverse Laplace transform of matrix exponential functions, and the anisotropic local law of the sample covariance matrix. The theoretical findings are validated on both synthetic data and the MNIST dataset using the Gaussian kernel and the neural tangent kernel.

Keywords

Kernel Gredient Descent, Spectral Bias, Neural Tangent Kernel, Wide Neural Network

1. 引言

在现代机器学习中，理解过参数化模型的泛化能力已成为核心理论问题之一。尤其是以深度神经网络为代表的非线性模型，尽管其参数规模通常远超训练样本数量，却依然能够在实际任务中取得良好的泛化性能。这一现象与经典统计学习理论中关于模型复杂度与过拟合之间的权衡关系形成了鲜明对比，因此引发了广泛关注与研究。

近年来，一条重要的研究路径是从函数空间而非参数空间的角度来分析学习算法的行为。特别地，在宽度趋于无穷的极限下，神经网络的训练动态可以用核方法进行刻画，其中最具代表性的结果是神经切线核(Neural Tangent Kernel, NTK)理论[1] [2]。该理论表明，在梯度下降训练过程中，神经网络的输出函数在函数空间中沿着某个核诱导的梯度方向演化，从而将非凸的参数优化问题转化为一个在函数空间中的线性动力系统。这一等价性建立了深度学习与核方法之间的桥梁，并带来了一系列关于神经正切核的研究[3]-[5]。

在核方法框架下，学习问题通常被建模为再生核希尔伯特空间(Reproducing Kernel Hilbert Space, RKHS) [6]中的函数估计问题。经典方法如核岭回归通过显式引入正则化项来控制模型复杂度，从而获得良好的泛化性能。然而，在实际应用中，许多基于梯度下降的算法并未显式加入正则化项，却依然能够避免过拟合并收敛到具有良好泛化能力的解。这一现象被称为隐式正则化(implicit regularization) [7] [8]，即优化算法本身在无显式约束的情况下偏向于某类“简单”或“低复杂度”的解。

已有研究表明，这种隐式偏置与模型的谱结构密切相关。例如，用梯度下降优化神经网络时的谱偏置(spectral bias)现象[9]，即学习算法倾向于优先拟合低频成分，而高频成分的学习则相对缓慢。这种现象在核岭回归的学习曲线分析中得到了严格刻画，泛化误差可以分解到协方差算子的不同特征子空间上，不同子空间的收敛速率由对应特征值决定[10] [11]。

尽管已有工作从上述多个角度揭示了隐式偏置的部分机制，但仍存在若干关键问题有待深入研究：首先，梯度下降在 RKHS 中的动态如何精确刻画；其次，这种动态如何在谱层面上体现为对不同特征模态的选择性学习；最后，这种选择性如何决定最终解的泛化能力。

基于上述背景，本文旨在系统研究再生核梯度下降(kernel gradient descent)优化过程中的谱偏置。在研究时我们采用了核梯度下降的连续时间形式——核梯度流。连续时间视角的一个优势在于，它让我们能够清楚地了解整个优化路径，而不仅仅像当前许多关于谱偏置的研究那样仅限于收敛点。我们从再生核希尔伯特空间的谱分解出发，分析核梯度流在不同特征子空间中的演化行为，并揭示训练过程如何通过时间演化实现对函数复杂度的有效控制。通过建立学习动态与核函数谱结构之间的定量关系，本文为理解过参数化模型的泛化能力提供了新的理论视角。

本文的主要贡献有：我们近似了核梯度下降在解决平方损失的回归问题时，其泛化误差整个优化过程中的变化，并且我们的结果将泛化误差分解为其在核的特征函数方向上的分量。实验表明，我们的结果在谱指数下降的高斯核以及谱幂律下降的神经正切核的核梯度下降中表现良好。

2. 核梯度流的泛化误差

2.1. 问题设定和记号

考虑核回归问题，其目标是在再生核希尔伯特空间中学习一个函数 $f:\mathcal{X}\to\mathbb{R}$ 来描述输入和输出间的关系。其中输入空间 $\mathcal{X}\subseteq\mathbb{R}^d$ 是紧的，输出空间 $\mathcal{Y}=\mathbb{R}$。设 ρ 是 $\mathcal{X}\times\mathcal{Y}$ 上的概率测度，$\{(x_i,y_i)\}_{i=1}^n\subset\mathcal{X}\times\mathcal{Y}$ 是从 ρ 中简单抽样得到的 n 组数据，ρ_X 是 ρ 在 $\mathcal{X}$ 上的边缘概率测度。记 $L^2=L^2(\mathcal{X},\mathrm{d}\rho_X)$ 为 $\mathcal{X}$ 上的平方可积函数类，$f^*(x)=\mathbb{E}_\rho[y\mid x]$ 为回归函数，并且 $\mathbb{E}_{(x,y)\sim\rho}\left[\left(y-f^*(x)\right)^2\mid x\right]=\sigma^2>0$，$\rho_X-\text{a.e. } x\in\mathcal{X}$。$\mathcal{H}$ 是以 $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ 为内积的再生核希尔伯特空间，对应的核函数为 $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$。我们希望在 $\mathcal{H}$ 中找到预测函数 f 使得下面的风险泛函尽可能小：

$$\mathcal{E}(f)=\mathbb{E}_{(x,y)\sim\rho}\left[\left(f(x)-y\right)^2\right]. \tag{1}$$

一般方法如核岭回归会最小化带二次正则项的经验风险：

$$\min_{f\in\mathcal{H}}\frac{1}{2n}\sum_{i=1}^n\left(f(x_i)-y_i\right)^2+\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2. \tag{2}$$

上式存在唯一解[12]：

$$\hat{f}(\boldsymbol{x})=\boldsymbol{y}^{\top}\left(\boldsymbol{K}+n\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{k}(\boldsymbol{x}), \tag{3}$$

其中 $\boldsymbol{K}$ 是 $n\times n$ 的 Gram 矩阵，$\boldsymbol{K}_{ij}=k(x_i,x_j)$，$\boldsymbol{k}(\boldsymbol{x})_i=k(\boldsymbol{x},\boldsymbol{x}_i)$。在本文中我们并未关注这种方法，因为其较大的计算复杂度 $O(n^3)$ 以及难以精准确定的正则化参数 λ，而是主要研究计算成本较低并且具有隐式正则化能力的核梯度下降[12]。

我们用过量风险来评估预测函数的表现，过量风险可以表示为预测函数和回归函数在 L^2 中距离的平方：

$$\forall f\in L^2,\ \mathcal{E}(f)-\mathcal{E}(f^*)=\left\|f^*-f\right\|_{L^2}^2. \tag{4}$$

2.2. 核梯度流

令式(2)中显示正则化强度 $\lambda=0$，并使用核梯度下降对其进行优化。在我们的理论研究中使用了核梯度下降的连续时间近似：核梯度流。核梯度下降和核梯度流将经验风险看作 $\mathcal{H}$ 上的泛函：

$$L(f)=\frac{1}{2n}\sum_{i=1}^n\left(f(x_i)-y_i\right)^2. \tag{5}$$

L 的梯度作为映射 $\nabla L:\mathcal{H}\to\mathcal{H}$ 由下式定义：

$$\left\langle \nabla L(f), g\right\rangle_{\mathcal{H}} = DL(f)(g),\ \forall g\in\mathcal{H}.$$

其中 $DL(f):\mathcal{H}\to\mathbb{R}$ 是 L 在 f 处的 Fréchet 导数。

核梯度流作为连续时间形式的核梯度下降，可以表示为以下的微分方程：

$$\frac{\partial f_t}{\partial t} = -\nabla L(f_t). \tag{6}$$

在本文的平方损失设定下，结合初值条件我们可以得到式(6)的具体形式：

$$\frac{\partial f_t}{\partial t} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_t(x_i)\right)k_{x_i},\ f_0 = 0. \tag{7}$$

上面的微分方程有唯一解：

$$f_t(x) = \boldsymbol{y}^{\top}\left(\boldsymbol{I} - \mathrm{e}^{-\frac{t}{n}\boldsymbol{K}}\right)\boldsymbol{K}^{-1}\boldsymbol{k}(x), \tag{8}$$

其中 $\boldsymbol{K}$ 是 $n\times n$ 的 Gram 矩阵，$\boldsymbol{K}_{ij} = k(x_i, x_j)$，且 $\boldsymbol{k}(x)_i = k(x, x_i)$ 与核岭回归时相同。这里我们假设 Gram 矩阵 $\boldsymbol{K}$ 是可逆的。

2.3. 泛化误差的谱分解表示

本文利用再生核希尔伯特空间的谱分解得到过量风险(4)的谱分解形式。再生核希尔伯特空间 $\mathcal{H}$ 对应的核函数 k 满足 $\sup_{x\in\mathcal{X}} k(x,x)<\infty$，从而有连续的嵌入 $\mathrm{id}:\mathcal{H}\to L^2$，令 $S_k: L^2\to\mathcal{H}$ 是其共轭算子，则称 $T = S_k^* S_k: L^2\to L^2$ 为协方差算子：

$$(Tf)(x) = \int_{\mathcal{X}} k(x,x')f(x')\mathrm{d}\rho_X(x'). \tag{9}$$

T 是一个正定的自伴紧算子[6]，由 Hilbert-Schmidt 定理，T 的所有特征向量构成 L^2 的一组正交基，进一步有 $L^2 = \ker(T)\overset{\perp}{\oplus} S$，其中 S 有至多可数的标准正交基 $\{\phi_i\}_{i\in I}$ 对应 T 的所有非零特征值 $\{\lambda_i\}_{i\in I}$，根据 Mercer 定理[6]：

$$k(x,x') = \sum_{i\in I}\lambda_i\phi_i(x)\phi_i(x'), \tag{10}$$

这里的收敛是绝对且一致的，并且 $\{\psi_i\}_{i\in I} = \{\sqrt{\lambda_i}\phi_i\}_{i\in I}$ 是 $\mathcal{H}$ 的一组标准正交基。

令测度 $\pi_m = \frac{1}{m}\sum_{i=0}^{m-1}\delta_{\lambda_i}$，其中 $\lambda_0\ge\lambda_1\ge\cdots\ge\lambda_m>0$ 是协方差算子 T 的前 m 个特征值，我们要求：
[labelwidth = 3em, leftmargin = 3em, labelsep = 0.5em]

(A1) 对任意 $m\ge 1$，存在足够小的 $\tau_m>0$，使得 $\lambda_0\le\tau_m^{-1}$，并且 $\pi_m([0,\tau])\le 1-\tau_m$。

上述条件来自[13]，它控制了 RKHS 谱的衰减速率不能过快使得几乎所有谱都在 0 附近。例如，所有以幂律衰减的谱均满足**(A1)**，但如果以指数速率 $\lambda_j\sim\beta^{-j}$ 衰减，则 $\beta<\tau_m^{-2/m\tau_m}$。

不同于经典的偏差–方差分解，我们利用紧算子的性质将泛化误差 $E_g(t) = \mathbb{E}\left[\left\|f^* - \hat{f}_t\right\|_{L^2}^2 \mid \boldsymbol{X}\right]$ 分解为协方差算子所有特征空间的模态误差之和。我们假设：[labelwidth = 3em, leftmargin = 3em, labelsep = 0.5em]

(A2) 回归函数 $f^*\in\mathcal{H}$，并且在 $\mathcal{H}$ 上有展开 $f^*(x) = \sum_{i\in I} w_i^*\psi_i(x)$。

同样预测函数 $\hat{f}$ 也能在 $\{\psi_i\}_{i\in I}$ 上展开：

$$\hat{f}_t(x)=\sum_{i\in I}\hat{w}_i(t)\psi_i(x). \tag{11}$$

I 是至多可数集 $|I|=M\le\infty$，令 $\boldsymbol{\Psi}\in\mathbb{R}^{M\times n}$ 为样本的特征矩阵，$\boldsymbol{\Psi}_{i,j}=\psi_i(x_j)$，并且：[labelwidth = 3em, leftmargin = 3em, labelsep = 0.5em]

(A3) 对任意 $p\in\mathbb{N}$，存在 C_p 使得 $\mathbb{E}_{\rho_X}|\boldsymbol{\Psi}_{i,j}|^p\le C_p$。

利用(8)以及再生核希尔伯特空间的再生性，可以得到 $\{\hat{w}_i(t)\}_{i\in I}$：

$$\hat{\boldsymbol{w}}(t)=\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}\right)^{-1}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}}\right)\boldsymbol{y}. \tag{12}$$

利用式(11)的表示方式我们可以将泛化误差 E_g 分解为不同特征值 λ_i 对应特征子空间的模态误差 $E_i(t)$：

$$E_g(t)=\sum_i E_i(t),\ E_i(t)=\lambda_i\mathbb{E}\left[\left(\hat{w}_i(t)-w_i^*\right)^2\mid\boldsymbol{X}\right]. \tag{13}$$

令 $\boldsymbol{\Lambda}$ 为协方差算子特征值构成的对角矩阵：$\boldsymbol{\Lambda}_{i,j}=\delta_{i,j}\lambda_i$，结合之前的计算，我们可以给出关于核梯度流泛化误差的第一个结果：

命题 2.1 在条件**(A1) (A2) (A3)**下，$\hat{\boldsymbol{w}}(t)$ 是核梯度流的解在不同特征模态的系数(12)，则其泛化误差 $E_g(t)$ 可以表示为：

$$E_g(t)=\mathrm{Tr}\left(\mathcal{R}(t)+\mathcal{N}(t)\right). \tag{14}$$

其中 $\mathcal{R}$ 是噪音无关项方阵，$\mathcal{N}$ 是噪音相关项方阵，每个对角元分别为预测函数在不同特征模态上的泛化误差，具体有：

$$\mathcal{R}(t)=\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}}\boldsymbol{\Lambda}\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}}\boldsymbol{w}^*\boldsymbol{w}^{*\top}, \tag{15}$$

$$\mathcal{N}(t)=\frac{\sigma^2}{n}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}}\right)\boldsymbol{\Lambda}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}}\right)\left(\frac{1}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\right)^{\dagger}. \tag{16}$$

其中 $\left(\frac{1}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\right)^{\dagger}$ 表示矩阵 $\frac{1}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$ 的 Moore-Penrose 广义逆。

利用矩阵指数函数的 Laplace 变换以及[10]和[14]中估计高斯过程回归的泛化误差的方法，我们可以计算式(15)和式(16)并得到：

命题 2.2 在条件**(A1) (A2) (A3)**下，令

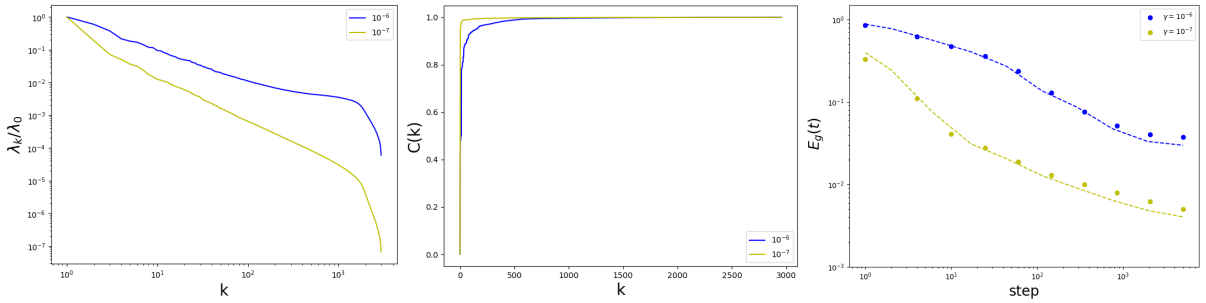
$$L_i(t)=\frac{1}{2\pi\mathrm{i}}\int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty}\frac{\mathrm{e}^{\frac{t}{n}s}}{s}\frac{s+T(s)}{s+T(s)+n\lambda_i}\mathrm{d}s, \tag{17}$$

其中 $\gamma>0$，$T(s)=\sum_i g_i(n,0)$ 是下面隐式方程的解：

$$T(s)=\sum_i\left(\frac{1}{\lambda_i}+\frac{n}{s+T(s)}\right)^{-1}. \tag{18}$$

则 $n\to\infty$ 时，核梯度流泛化误差的特征模态分解(13)式 $E_g(t)=\sum_i E_i(t)$ 中 $E_i(t)$ 可表示为：

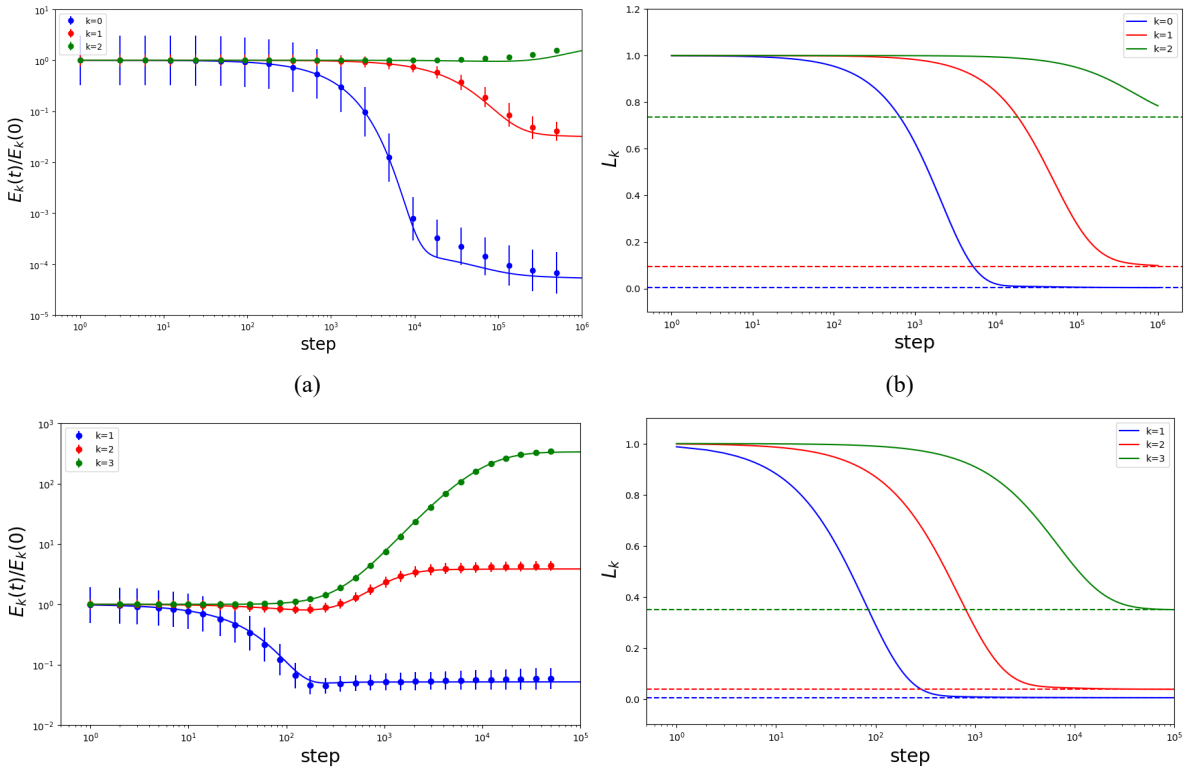
$$E_i(t)=\left(1+o_{\mathbb{P}}(1)\right)\left(\lambda_i L_i(t)^2 w_i^{*2}+\frac{\sigma^2}{n}\left(1-L_i(t)\right)^2\right). \tag{19}$$



(a) (b) (c)

Figure 1. Using kernel gradient descent with Gaussian kernels $k(x,x')=\mathrm{e}^{-\gamma\|x-x'\|^2}$ under different bandwidth parameters γ to learn the classification of digits 0 and 1 from the MNIST dataset: (a) Spectra of Gaussian kernels with two bandwidths; (b) $C(k)$ of Gaussian kernels with two bandwidths; (c) Generalization errors of kernel gradient descent for Gaussian kernels with two bandwidths

图 1. 使用不同带宽参数 γ 的高斯核 $k(x,x')=\mathrm{e}^{-\gamma\|x-x'\|^2}$ 核梯度下降学习 MNIST 数据集的 0、1 手写数字分类问题。(a) 两种带宽的高斯核的谱。(b) 两种带宽的高斯核的 $C(k)$。(c) 两种带宽的高斯核核梯度下降的泛化误差



(a) (b) (c) (d)

Figure 2. (a) Modal generalization error of kernel gradient descent with Gaussian kernel as a function of iteration step. The solid line represents the estimate given by our results, and the error bars are from simulation experiments repeated 10 times. (b) Modal residual L_i from the same experiment. (c) Modal generalization error of kernel gradient descent with neural tangent kernel as a function of iteration step. (d) Modal residual L_i from the same experiment

图 2. (a) 使用高斯核进行核梯度下降的模态泛化误差随迭代步的变化。实线为我们的结果给出的估计，误差线为重复 10 次的模拟实验。(b) 相同实验的模态剩余学习程度 L_i。(c) 使用神经正切核进行核梯度下降的模态泛化误差随迭代步的变化。(d) 相同实验的模态剩余学习程度 L_i

注意到，每个模态的 $E_i(t)$ 都可分解为 σ 无关项和与 σ 相关项两部分，后者完全刻画了噪声对泛化误差的影响。

式(19)中的 $L_i(t)$ 可以理解为核梯度流对回归函数分模态的剩余学习程度。对任意模态 i，在 $t=0$ 时，$L_i(t)=1$，并在核梯度流过程中单调减少，且有 $\lim_{t\to\infty} L_i(t)=T(0)/(T(0)+n\lambda_i)$。同时，根据我们的实验，较大特征值对应的模态往往学习速度较快，表现为对应的 $L_i(t)$ 下降得较快。上述推论直接导致了：

$$\lambda_i > \lambda_{i'} \Rightarrow L_i < L_{i'}. \tag{20}$$

这种现象被称为谱偏置[9]，即核梯度下降对目标函数的谱分量有不同的学习倾向。谱偏置最初在神经网络的研究中被提出，这一概念颇具吸引力，因为它直观地解释了为何过参数化的神经网络能在不发生过拟合的情况下取得良好的泛化性能：在训练过程中，网络先拟合目标函数在频率较低的模态上的分量，所以泛化效果更好。随后产生了一系列对神经正切核的谱偏置的研究[3] [4]。

同时从我们的结果也能看出，核梯度下降对样本的学习不仅会学到真实的模态系数 w_i^*，也会学习到噪音，后者直接导致了过拟合。在式(19)中体现为：随着核梯度下降的进行，$L_i(t)$ 减小，第一项表示对真实回归函数的学习，所以使得泛化误差减小，而第二项表示对噪声的学习，这会导致泛化误差的增加。在实验中画出 $E_i(t)/E_i(0)$ 的学习曲线，我们通常可以观察到两种不同的变化趋势，分别是单调下降和先下降后上升。对式(19)关于时间 t 求导可以得到：$E_i'(t)=2L_i'(t)\left(\lambda_i w_i^{*2}L_i(t)-\frac{\sigma^2}{n}(1-L_i(t))\right)$，第一个乘项 $L_i'(t)<0$，所以我们关注第二个乘项，它决定了学习曲线是否单调下降，其零点在满足 $L_i(t)=\frac{\sigma^2}{n}\Big/\left(\lambda_i w_i^{*2}+\frac{\sigma^2}{n}\right)$ 的 t 处，我们称 $SNR_i=n\lambda_i w_i^{*2}/\sigma^2$ 为模态信噪比。当模态信噪比较大使得 $\frac{1}{SNR_i+1}<\frac{T(0)}{T(0)+n\lambda_i}$ 时，$E_i(t)/E_i(0)$ 单调下降，否则 $E_i(t)/E_i(0)$ 会先下降后上升。

核函数的选择以及核函数中超参数的选择是核方法的关键，而式(19)中的第一项可以帮助我们理解什么样的核函数适合当前的问题。令分布

$$C(\rho)=\frac{\sum_{i\le\rho}\lambda_i w_i^{*2}}{\sum_i \lambda_i w_i^{*2}}, \tag{21}$$

由式(20)我们注意到，对同一个学习问题，选择的核函数使 $C(\rho)$ 越大，则式(19)中的第一项越小，称为任务与模型的对齐[11]，曲线 $C(\rho)$ 的下方面积可以用来作为任务模型对齐程度的度量。

图 1 中我们使用不同带宽的高斯核进行 MNIST 数据集[15]中 0 和 1 的分类。我们将该任务建模为一个核回归问题，具体方法是考虑一个向量值的目标函数，该函数以手写数字图片为输入，输出独热编码标签。我们的核梯度下降理论可以分别应用于输出的每个分量，并且可以通过将每个分量产生的误差相加来计算总的泛化误差。在下一小节我们介绍了具体的做法。由图 1 可知，不同的带宽参数使得高斯核的谱衰减速率不同，进而与任务的对齐程度不同，较高的 $C(\rho)$ 与任务对齐得更好，所以泛化误差更低。

2.4. 在真实数据集上应用

计算 E_g 需要两个输入：核特征值 λ_i 和教师权重 w_i^*，两者都需要使用样本真实分布计算。对于有限样本的数据集，我们用数据的经验分布 $p(x)=\frac{1}{n}\sum_{i=1}^n \delta(x-x_i)$ 近似真实分布，于是相应的特征值问题为：

$$\lambda_k \phi_k(x')=\int p(x)k(x,x')\phi_k(x)\mathrm{d}x=\frac{1}{n}\sum_{i=1}^n k(x_i,x')\phi_k(x_i).$$

设 $\boldsymbol{K}$ 是核 Gram 矩阵 $\boldsymbol{K}=k\left(x_i,x_j\right)$，此时求核函数的特征值与特征函数相当于求解一个核主成分问题，即求 $\boldsymbol{K}=n\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{\top}$ 的特征值 $\boldsymbol{\Lambda}_{kl}=\lambda_k\delta_{kl}$ 和特征函数 $\boldsymbol{\Phi}_{ki}=\phi_k\left(x_i\right)$。

对于具有多类别的目标函数 $f^*\left(x\right):\mathbb{R}^D\to\mathbb{R}^C$，输出样本为独热编码 $\boldsymbol{y}=\left[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_C\right]\in\mathbb{R}^{n\times C}$，通过 $\boldsymbol{w}_c^*=\frac{1}{n}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Phi}^{\top}\boldsymbol{y}_c$ 得到目标函数在每个模态的权重。对于核梯度下降及其泛化误差，每个输出分量都可看成是一个标量函数，我们对每个分量进行核梯度下降并且对每个分量的泛化误差求和得到总的泛化误差。

3. 模拟实验

在本节中，我们分别通过核梯度下降和宽神经网络模拟实验来验证我们的理论结果。在样本输入服从高斯分布时，高斯核的谱指数衰减[16]，在样本服从超球面上的均匀分布时，神经正切核的谱幂律衰减[5]。我们分别使用这两种不同谱衰减速率的核进行核梯度下降。

3.1. 高斯核核梯度下降

我们首先使用高斯核验证我们的理论，即 $k\left(x,x'\right)=\mathrm{e}^{-\frac{1}{2\ell^2}\left\|x-x'\right\|^2}$，当输入分布为 $\mathbb{R}^d$ 上每个分量独立等方差的多元正态分布时，其积分算子的特征值 $\left\{\lambda_k\right\}$ 和特征函数可以显式计算[16]。在 $d=1$ 时：

$$\lambda_k=\sqrt{\frac{2a}{A}}B^k, \tag{22}$$

$$\phi_k\left(x\right)=R_k\sqrt{\frac{c}{a}}\exp\left(-\left(c-a\right)x^2\right)H_k\left(\sqrt{2c}x\right), \tag{23}$$

其中 $H_k\left(x\right)=\left(-1\right)^k\exp\left(x^2\right)\frac{\mathrm{d}^k}{\mathrm{d}x^k}\exp\left(-x^2\right)$ 是 k 阶 Hermite 多项式，$R_k=\frac{1}{2^k k!}$ 是归 $t\to\infty$ 一化常数，$a^{-1}=4\sigma^2$，$b^{-1}=2\ell^2$，且

$$c=\sqrt{a^2+2ab},\ A=a+b+c,\ B=b/A. \tag{24}$$

在多维情形下，核函数和高斯密度是单变量的乘积形式，特征值和特征函数的结果可以很容易地推广到多维情形，此时特征函数和特征值也仅仅是乘积形式。对于所有 d 个维度上 a 和 b 都相等的情况，特征值 $\left(\frac{2a}{A}\right)^{d/2}B^k$ 的重数 $N\left(d,k\right)$ 为 $\binom{k+d-1}{d-1}\sim\mathcal{O}\left(k^{d-1}\right)$，特征值大小呈指数衰减。

我们用高斯核的核梯度下降验证命题 2.2 中的估计。在实验中，我们随机抽取一系列点 $\left\{\bar{x}_i\right\}$，并将 $\left\{k_{\bar{x}_i}\right\}$ 的线性组合作为回归函数：

$$f^*\left(x\right)=\sum_{i=1}^{n'}\bar{\alpha}_i k\left(x,\bar{x}_i\right). \tag{25}$$

这里 $\bar{\alpha}_i\sim\mathcal{B}\left(1/2\right)$ 是在 $\left\{\pm1\right\}$ 中取值的伯努利分布，$\bar{x}_i$ 和样本分布相同，且与训练样本独立。实验中用离散形式的核梯度下降[8]来实现式(7)中的核梯度流[17]：

$$\begin{aligned}f_t&=f_{t-1}+\gamma\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\left\langle f_{t-1},k_{x_i}\right\rangle_{\mathcal{H}}\right)k_{x_i}\\&=f_{t-1}+\gamma\frac{1}{n}\sum_{i=1}^{n}\left(y_i-f_{t-1}\left(x_i\right)\right)k_{x_i},\end{aligned} \tag{26}$$

其中 γ 是常数步长。从上面的迭代中可以看出 f_t 可表示为基核函数 $\left\{k_{x_i}\right\}$ 的线性组合，$f_t=\alpha\left(t\right)^{\top}\boldsymbol{k}\left(x\right)$，则 α 的迭代为：

$$\alpha_t = \alpha_{t-1} + \gamma \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{K}\alpha_{t-1}\right),\ \alpha_0 = 0. \tag{27}$$

在此设定下，模态误差 E_k 可再次被分解为 $E_k = \sum_m^{N(d,k)} E_{km}$，我们将这些实验中的模态误差与理论预测值进行对比。

在图 2(a)中，我们将 $n = 100$ 个 $d = 20$ 维标准正态随机向量作为输入样本，噪声方差 $\sigma^2 = 0.25$。我们使用带宽参数 $\ell = 5$ 的高斯核进行核梯度下降，并取前三个模态的泛化误差与我们的理论估计进行了比较，我们的理论良好地预测了实验结果。在图 2(b)中，实线是优化过程中对应模态的 $L_i(t)$，虚线是时的 $L_i(t)$ 的极限。

3.2. 神经正切核核梯度下降

全连接神经网络的训练演化过程可由核函数描述[1]：在对网络参数 θ 进行梯度下降时，网络的输出函数 f_θ 遵循损失泛函关于神经正切核的核梯度下降。神经正切核对于描述神经网络的泛化特性至关重要。尽管其在初始化时具有随机性且在训练过程中是变化的，但在参数正态初始化以及 Lipschitz 连续的 ReLU 激活设定下，当 L 层网络的层宽 $n_0, \cdots, n_L$ 趋于无穷时，神经正切核将依概率收敛于一个非随机的极限核且在训练中保持恒定[1]。在后文中为了方便，我们提到的神经正切核均指其极限核。在神经正切核上应用我们的结果可以得到无限宽全连接神经网络的泛化误差优化曲线。

下面我们先用神经正切核的核梯度下降验证命题 2.2。回归函数的形式与式(25)相同，$\bar{\alpha}_i \sim \mathcal{B}(1/2)$ 是在 $\{\pm 1\}$ 中取值的伯努利分布，$\bar{x}_i$ 服从单位超球面 $\mathbb{S}^{d-1}$ 上的均匀分布，且与训练样本独立。

通过单位超球面上点积核的分解可以得到神经正切核的 Mercer 分解。我们考虑输入空间 $\mathcal{X} = \mathbb{S}^{d-1}$ 为 d 维单位超球面，ρ_X 为 $\mathbb{S}^{d-1}$ 上的均匀分布。此时，形式为 $k(x, x') = \kappa\left(x^\top x'\right)$ 的点积核具有旋转不变性，其取值仅依赖于 x 和 x' 之间的夹角，并且其协方差算子 T 在球谐函数基下对角化[18]，即 $TY_{k,j} = \lambda_k Y_{k,j}$，其中 $Y_{k,j}$ 是 k 阶的第 j 个球谐多项式，而同一特征值 λ_k 的重数为 $N(d,k) = \dfrac{2k+d-2}{k}\binom{k+d-3}{d-2}$，这意味着 k 阶正交的球谐多项式有 $N(d,k)$ 个，当 k 很大时，它以 k^{d-2} 的速度增长。特征值 λ_k 仅依赖于阶数 k，并由下式给出[19]：

$$\lambda_k = \frac{\omega_{d-1}}{\omega_d}\frac{1}{N(d,k)}\frac{k+\alpha}{\alpha}\int_{-1}^{1}\kappa(t)Q_k^\alpha(t)\left(1-t^2\right)^{\frac{d-3}{2}}\mathrm{d}t, \tag{28}$$

其中 $\alpha = \dfrac{d}{2} - 1$，而 Q_k^α 是 d 维空间中的 k 次 Gegenbauer 多项式[19]，ω_d 表示球面 S^{d-1} 的表面积。由 Mercer 定理：

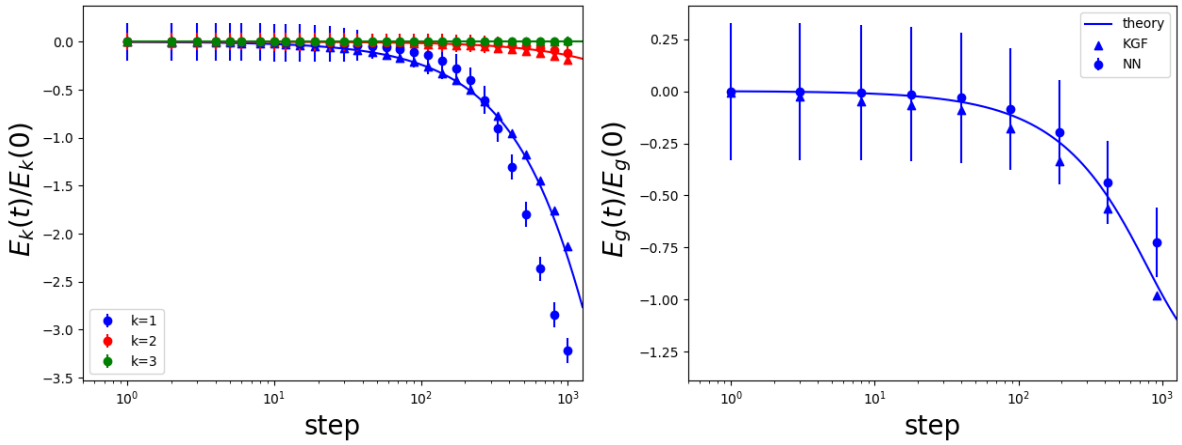
$$k(x,x') = \sum_{k=0}^{\infty}\lambda_k \sum_{j=1}^{N(d,k)} Y_{kj}(x)Y_{kj}(x'). \tag{29}$$

在数值模拟中，我们通过 Gauss-Gegenbauer 求积法[20]数值计算式(28)得到极限神经正切核的特征值。

我们将对比实验中的经验模态误差 $E_k = \sum_j^{N(d,k)} E_{kj}$ 和我们的理论结果。在这些实验中，k_{NTK} 均为无偏置的三层全连接 ReLU 网络对应的极限神经正切核。

在图 2(c)中，我们将 $n = 1000$ 个 $d = 10$ 维 $\mathbb{S}^{d-1}$ 上均匀分布输入样本的神经正切核核梯度下降模态泛化误差与我们的理论估计进行了比较，噪音方差为 $\sigma^2 = 0.1$，我们的理论较好的预测了实验结果。在图 2(d)中，实线是优化过程中对应模态的 $L_i(t)$，虚线是 $t \to \infty$ 时的 $L_i(t)$ 的极限。对比两种不同衰减速度的核函数我们发现，谱指数衰减的高斯核，模型几乎只在极低频的子空间内工作，天然地过滤掉了高频成

分。而谱幂律衰减的神经正切核，它的特征值分布厚尾。这意味着即使是高频模态，仍然分配到了不可忽视的权重。这虽然增强了它的表达能力，但也让它对噪声更加敏感。可以看到即使在 NTK 核的实验中噪音方差更小，学习曲线还是因为噪音产生了拐点，主要原因是，衰减缓慢的谱的 $T(0)$ 较大。注意到在式(25)的设定下，$\mathbb{E}w_i^{*2}=\lambda_i$，于是要使 $\frac{1}{SNR_i+1}>\frac{T(0)}{T(0)+n\lambda_i}$，只需 $T(0)>\lambda_i/\sigma^2$。



(a) (b)

Figure 3. Generalization error of a fully connected ReLU neural network as a function of iteration step. The solid line represents the estimate given by our results, the triangles represent the results of kernel gradient descent, the circles represent the results of the neural network, and the error bars are from simulation experiments repeated 10 times. (a) Generalization error of the corresponding mode under the single-mode setting. (b) Total generalization error under the full-mode setting

图 3. 使用全连接 ReLU 神经网络泛化误差随迭代步的变化。实线为我们的结果给出的估计，三角形为核梯度下降的结果，圆形为神经网络的结果，误差线为重复 10 次的模拟实验。(a) 单一模态设定下，对应模态的泛化误差。 (b)全模态设定下，总泛化误差

3.3. 有限宽神经网络

在证实了我们的理论能够准确预测神经正切核核梯度下降的泛化误差之后，我们现在将有限宽神经网络在平方损失函数下训练得到的泛化误差，与神经正切核的理论学习曲线进行比较。

首先，我们采用单一模态的回归函数，即回归函数仅由相同阶数的球谐函数构成。对于阶数为 k 的模态，回归函数如下构建：

$$f^*(x)=\sum_{i=1}^{n'}\bar{\alpha}_i Q_k^{d/2-1}\left(x^\top \bar{x}_i\right), \tag{30}$$

其中 $\bar{\alpha}_i\sim\mathcal{B}(1/2)$，$\bar{x}_i\sim\rho_X(x)$ 均为随机采样。图 3(a)展示了宽度 $M=5000$、输入维度 $d=10$ 且 $n'=10000$ 的全连接两层 ReLU 网络的学习曲线。与前文相同的，我们观察到阶数 k 较低的模态学习得更快。

接着，我们证明我们的理论同样适用于包含多种不同阶球谐函数的复合函数。在此实验中，我们随机初始化一个两层神经网络作为回归函数，并训练一个神经网络学习它：

$$f^*(x)=\bar{r}^\top\sigma\left(\bar{\Theta}x\right),\ f(x)=r^\top\sigma(\Theta x). \tag{31}$$

其中 $\Theta,\bar{\Theta}\in\mathbb{R}^{M\times d}$ 分别为学习函数和回归函数的前馈权重，σ 为激活函数，$r,\bar{r}\in\mathbb{R}^M$ 为学习函数和回归函数的读出权重。采用 ReLU 激活函数如此构建时，回归函数由多种不同阶数的球谐函数组成。对于 $d=10$、$M=5000$ 的超参数设置，图 3(b)展示了该情形下的总泛化误差以及我们理论的预测值。

4. 结论

本文给出了用核梯度下降优化最小二乘回归问题时，泛化误差在优化过程中的变化的一个近似表达式，通过该表达式我们可以理解核函数的选择是怎样影响泛化效果的，具体表现为其谱的衰减速度以及任务模型的对齐。我们分别在正态数据的高斯核以及单位超球面上均匀分布数据的神经正切核上验证了我们的估计，它们的谱分别呈指数衰减和幂律衰减，结果表现良好。我们进一步用高斯核在真实数据集MNIST上进行实验并验证了任务模型的对其理论。最后，我们用宽神经网络验证我们的理论，但由于宽神经网络的初始输出函数非零，而与我们的结果产生偏差，这也是我们理论需要完善的方向。

5. 补充证明

式(8)的证明。由式(7)可知，在任意时间t，f_t可以表示为$\mathcal{H}$中函数集$\left\{k_{x_i}\right\}_{i=1}^n$的线性组合。于是可令$f_t(x)=\alpha(t)^\top \boldsymbol{k}(x)$，我们只需关注$t$时刻在样本上的预测$f_t(\boldsymbol{x})=\boldsymbol{K}\alpha(t)$，于是有式(7)对应的参数形式常微分方程组：

$$\frac{\mathrm{d}\alpha(t)}{\mathrm{d}t}=\frac{1}{n}\left(\boldsymbol{y}-\boldsymbol{K}\alpha(t)\right),\ \alpha(0)=0. \tag{32}$$

令$\eta(t)=\boldsymbol{K}\alpha(t)-\boldsymbol{y}$，我们得到：

$$\frac{\mathrm{d}\eta(t)}{\mathrm{d}t}=\boldsymbol{K}\frac{\mathrm{d}\alpha(t)}{\mathrm{d}t},$$

且方程(32)可写为：

$$\frac{\mathrm{d}\eta(t)}{\mathrm{d}t}=-\frac{1}{n}\boldsymbol{K}\eta(t),\ \eta_0=-\boldsymbol{y}$$

解得：

$$\eta(t)=-\exp\left(-\frac{t}{n}K\right)\boldsymbol{y}.$$

代回原式我们得到：

$$\alpha(t)=\boldsymbol{K}^{-1}\left(\boldsymbol{I}-\exp\left(-\frac{t}{n}\boldsymbol{K}\right)\right)\boldsymbol{y}, \tag{33}$$

代回到$f_t(x)=\alpha(t)^\top \boldsymbol{k}(x)$即得到式(8)。

命题 2.1 的证明。由式(13)，

$$E_g(t)=\mathbb{E}\left[\left(\hat{\boldsymbol{w}}(t)-\boldsymbol{w}^*\right)^\top \boldsymbol{\Lambda}\left(\hat{\boldsymbol{w}}(t)-\boldsymbol{w}^*\right)\mid \boldsymbol{X}\right],$$

其中：

$$\begin{aligned}\hat{\boldsymbol{w}}(t)&=\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)^{-1}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right)\boldsymbol{y}\\&=\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)^{-1}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right)\left(\boldsymbol{\Psi}^\top\boldsymbol{w}^*+\boldsymbol{\varepsilon}\right)\\&=\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right)\boldsymbol{w}^*+\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right)\boldsymbol{\Psi}^{\dagger\top}\boldsymbol{\varepsilon}\end{aligned}$$

于是利用迹与期望的可交换性得到：

$$E_g(t)=\mathrm{Tr}\left(\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\boldsymbol{\Lambda}\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\boldsymbol{w}^*\boldsymbol{w}^{*\top}+\frac{\sigma^2}{n}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right)\boldsymbol{\Lambda}\left(\boldsymbol{I}-\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right)\left(\frac{1}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right)^\dagger\right)$$

命题 2.2 的证明。首先，我们用矩阵指数函数的 Laplace 逆变换表示：

$$\mathrm{e}^{-At}=\frac{1}{2\pi\mathrm{i}}\int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty}\mathrm{e}^{-st}\left(s\boldsymbol{I}+A\right)^{-1}\mathrm{d}s, \tag{34}$$

其中 γ 大于 $-A$ 的所有特征值实部的最大值。将 $A=\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ 带入到式(34)中，时间尺度定为 t/n，接着对两边关于样本取期望有：

$$\mathbb{E}\left[\mathrm{e}^{-\frac{t}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top}\right]=\frac{1}{2\pi\mathrm{i}}\int_{\gamma-\mathrm{i}\infty}^{\gamma+\mathrm{i}\infty}\mathrm{e}^{\frac{t}{n}s}\mathbb{E}\left[\left(s\boldsymbol{I}+\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right)^{-1}\right]\mathrm{d}s. \tag{35}$$

上式中 $\mathbb{E}\left[\left(s\boldsymbol{I}+\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right)^{-1}\right]$ 又可进一步表示为 $\frac{1}{s}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbb{E}\left[\left(\boldsymbol{\Lambda}^{-1}+\frac{1}{s}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\right)^{-1}\right]\boldsymbol{\Lambda}^{-\frac{1}{2}}$，其中 $\boldsymbol{\Phi}=\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Psi}$。

我们引入一个辅助变量 v 并令：

$$\boldsymbol{G}(n,v)=\left(\frac{1}{s}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top+\boldsymbol{\Lambda}^{-1}+v\boldsymbol{I}\right)^{-1}. \tag{36}$$

考虑添加一个从总体分布中随机抽取的新输入样本 $x'\sim\rho_X$ 对 $\boldsymbol{G}$ 的影响，令 $\boldsymbol{\phi}\in\mathbb{R}^M$ 为该新输入样本的特征，即 $\phi_i=\phi_i(x')$，$\boldsymbol{X}$ 则代表已有输入样本。由 Woodbury 逆矩阵公式：

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X},x'}\boldsymbol{G}(n+1,v)&=\mathbb{E}_{\boldsymbol{X},x'}\left(\boldsymbol{G}(n,v)^{-1}+\frac{1}{s}\boldsymbol{\phi}\boldsymbol{\phi}^\top\right)^{-1}\\&=\mathbb{E}_{\boldsymbol{X}}\boldsymbol{G}(n,v)-\mathbb{E}_{\boldsymbol{X},x'}\frac{\boldsymbol{G}(p,v)\boldsymbol{\phi}\boldsymbol{\phi}^\top\boldsymbol{G}(p,v)}{s+\boldsymbol{\phi}^\top\boldsymbol{G}(p,v)\boldsymbol{\phi}}.\end{aligned} \tag{37}$$

对上式右边最后一项取期望是困难的，因此我们采用一种近似方法，即对分子和分母分别取期望：

$$\mathbb{E}_{\boldsymbol{X},x'}\boldsymbol{G}(n+1,v)\approx\mathbb{E}_{\boldsymbol{X}}\boldsymbol{G}(p,v)-\frac{\mathbb{E}_{\boldsymbol{X}}\boldsymbol{G}(n,v)^2}{s+\mathrm{Tr}\mathbb{E}_{\boldsymbol{X}}\boldsymbol{G}(n,v)}. \tag{38}$$

接着将 n 视为连续的变量，并利用 $\mathbb{E}\boldsymbol{G}^2(n,v)=-\frac{\partial}{\partial v}\mathbb{E}\boldsymbol{G}(n,v)$ 最终得到一个偏微分方程：

$$\frac{\partial\mathbb{E}\boldsymbol{G}(n,v)}{\partial n}=\frac{1}{s+\mathrm{Tr}\mathbb{E}\boldsymbol{G}(n,v)}\frac{\partial}{\partial v}\mathbb{E}\boldsymbol{G}(p,v). \tag{39}$$

注意到初值条件 $\mathbb{E}\boldsymbol{G}(n,v)=\left(\boldsymbol{\Lambda}^{-1}+v\boldsymbol{I}\right)^{-1}$，结合式(38)可以看出对任意 n 和 v，$\mathbb{E}\boldsymbol{G}$ 都是对角阵。求解该微分方程，我们可以得到矩阵 $\mathbb{E}\boldsymbol{G}(n,v)$ 的对角线元 $g_i(n,v)=\mathbb{E}\boldsymbol{G}(n,v)_{ii}$ 的表示：

$$g_i(n,v)=\left(\frac{1}{\lambda_i}+v+\frac{n}{s+\sum_{\gamma=1}^{M}g_\gamma(n,v)}\right)^{-1}. \tag{40}$$

接着我们用[13]中关于样本协方差矩阵各向异性局部律的推论 3.9 来证明(19)，在条件**(A1)(A2)(A3)**下有：

$$\boldsymbol{\Lambda}^{-1/2}\left(R_M(z)-\frac{-1}{z\left(1+m(z)\boldsymbol{\Lambda}\right)}\right)\boldsymbol{\Lambda}^{-1/2}=O_{\mathbb{P}}\left(\eta(z)\right).$$

其中 $R_M(z)=\left(\frac{1}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}-z\boldsymbol{I}_M\right)^{-1}$ 是 $\frac{1}{n}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}$ 的预解式，$\eta(z)=\sqrt{\frac{\operatorname{Im}m(z)}{n\operatorname{Im}z}}+\frac{1}{n\operatorname{Im}z}$，$m(z)$ 满足隐式方程 $\frac{1}{m(z)}=-z+\frac{1}{n}\sum_i\frac{\lambda_i}{1+m(z)\lambda_i}$，代入 Laplce 逆变换中可得(19)。

参考文献

[1] Jacot, A., Gabriel, F. and Hongler, C. (2018) Neural Tangent Kernel: Convergence and Generalization in Neural Networks. 32*nd International Conference on Neural Information Processing Systems* (*NeurIPS* 2018), Montreal, 3-8 December 2018, 8580-8589.

[2] Arora, S., Du, S.S., Hu, W., Li, Z., Salakhutdinov, R.R. and Wang, R. (2019) On Exact Computation with an Infinitely wide Neural Net. 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 8141-8150.

[3] Bietti, A. and Mairal, J. (2019) On the Inductive Bias of Neural Tangent Kernels. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 12893-12904.

[4] Cao, Y., Fang, Z., Wu, Y., Zhou, D. and Gu, Q. (2021) Towards Understanding the Spectral Bias of Deep Learning. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, 21-26 August 2021, 2205-2211. https://doi.org/10.24963/ijcai.2021/304

[5] Bietti, A. and Bach, F. (2021) Deep Equals Shallow for ReLU Networks in Kernel Regimes. *International Conference on Learning Representations*, Vienna, 4 May 2021, 12913-12934.

[6] Steinwart, I. and Christmann, A. (2008) Support Vector Machines. Springer.

[7] Ali, A., Dobriban, E. and Tibshirani, R. (2020) The Implicit Regularization of Stochastic Gradient Flow for Least Squares. *International Conference on Machine Learning*, Vienna, 13-18 July 2020, 233-244.

[8] Yao, Y., Rosasco, L. and Caponnetto, A. (2007) On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, **26**, 289-315. https://doi.org/10.1007/s00365-006-0663-2

[9] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y. and Courville, A. (2019) On the Spectral Bias of Neural Networks. *Proceedings of the* 36*th International Conference on Machine Learning*, Long Beach, 9-15 June 2019, 5301-5310.

[10] Bordelon, B., Canatar, A. and Pehlevan, C. (2020) Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. *Proceedings of the* 37*th International Conference on Machine Learning*, Vienna, 13-18 July 2020, 1024-1034.

[11] Canatar, A., Bordelon, B. and Pehlevan, C. (2021) Spectral Bias and Task-Model Alignment Explain Generalization in Kernel Regression and Infinitely Wide Neural Networks. *Nature Communications*, **12**, Article No. 2914. https://doi.org/10.1038/s41467-021-23103-1

[12] Allerbo, O. (2025) Fast Robust Kernel Regression through Sign Gradient Descent with Early Stopping. *Electronic Journal of Statistics*, **19**, 1231-1285. https://doi.org/10.1214/25-ejs2361

[13] Knowles, A. and Yin, J. (2016) Anisotropic Local Laws for Random Matrices. *Probability Theory and Related Fields*, **169**, 257-352. https://doi.org/10.1007/s00440-016-0730-4

[14] Sollich, P. (1998) Learning Curves for Gaussian Processes. *Proceedings of the* 12*th International Conference on Neural Information Processing Systems*, Denver, 1-3 December 1998, 344-350.

[15] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[16] Rasmussen, C.E. and Williams, C.K.I. (2005) Gaussian Processes for Machine Learning. The MIT Press. https://doi.org/10.7551/mitpress/3206.001.0001

[17] Li, Y., Gan, W., Shi, Z. and Lin, Q. (2024) Generalization Error Curves for Analytic Spectral Algorithms under Power-law Decay. arXiv:2401.01599.

[18] Smola, A., Óvári, Z. and Williamson, R.C. (2000) Regularization with Dot-Product Kernels. *Proceedings of the* 14*th International Conference on Neural Information Processing Systems*, Denver, 1 January 2000, 290-296.

[19] Dai, F. and Xu, Y. (2013) Approximation Theory and Harmonic Analysis on Spheres and Balls. Springer.

[20] Ralston, A. and Rabinowitz, P. (2001) A First Course in Numerical Analysis. Dover Publications.