Automated Spectral Kernel Learning

Jian Li, Yong Liu* and Weiping Wang
{lijian9026, liuyong, wangweiping}@iie.ac.cn

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Introduction

The generalization performance of kernel methods is largely determined by the kernel, but spectral representations of stationary kernels are both input-independent and output-independent, which limits their applications on complicated tasks.

In this paper, to achieve better performance ability for kernel methods, we propose an efficient algorithm, namely Automated Spectral Kernel Learning (ASKL), learning suitable kernels and model weights together.

1. Core Idea: Spectral kernel representations + Rademacher complexity.

2. Algorithmic contributions:

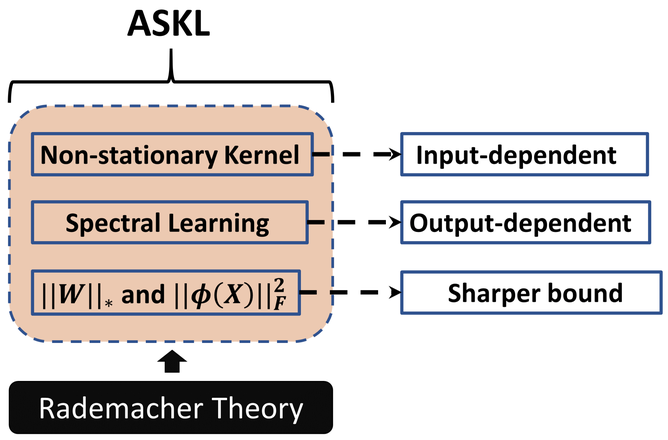
(1) non-stationary kernels to obtain input-dependent features

(2) backpropagation w.r.t the objective to make features output-dependent

(3) regularization terms to achieve sharper generalization error bounds

3. Theoretical contributions:

Based on **Rademacher complexity theory**, we explore how the feature mappings affect the performance and suggests ways to refine the algorithm.



Problem Definition

In ordinary supervised learning settings, training samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^n\}$ are drawn i.i.d. from a fixed but unknown distribution ρ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^d$ is the input space and $\mathcal{Y} \subseteq \mathbb{R}^K$ is the output space in single-valued ($K = 1$) or vector-valued ($K > 1$) forms. The goal is to learn an estimator $f : \mathcal{X} \to \mathcal{Y}$, which outputs K predictive labels. We define a standard hypothesis space for kernel methods

$$\mathcal{H} = \left\{ \boldsymbol{x} \to f(\boldsymbol{x}) = \boldsymbol{W}^T \phi(\boldsymbol{x}) \right\},$$

where $\boldsymbol{W} \in \mathbb{R}^{D \times K}$ is the model weight, $\phi(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^D$ is a nonlinear feature mapping. For kernel methods, $\phi(\boldsymbol{x})$ is an implicit feature mapping associated with a Mercer kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$. To improve the computational efficiency but also retain favorable statistical properties, random Fourier features were proposed to approximate kernel with explicit feature mappings $\phi(\boldsymbol{x})$ via $k(\boldsymbol{x}, \boldsymbol{x}') \approx \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$.

In statistical learning theory, the supervised learning problem is to minimize the expected loss on $\mathcal{X} \times \mathcal{Y}$

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

where ℓ is a loss function related to specific tasks.

Learning Framework

Based on Yaglom's theorem, the nonlinear feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^D$ adopts Monte Carlo approximation via inverse Fourier transform

$$\begin{aligned}
& k(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{E}_{\boldsymbol{\omega}, \boldsymbol{\omega}'}(\boldsymbol{x}, \boldsymbol{x}') \mu(d\boldsymbol{\omega}, d\boldsymbol{\omega}') \\
= \; & \mathbb{E}_{\boldsymbol{\omega}, \boldsymbol{\omega}' \sim s} \left[\mathcal{E}_{\boldsymbol{\omega}, \boldsymbol{\omega}'}(\boldsymbol{x}, \boldsymbol{x}') \right] \\
= \; & \mathbb{E}_{\boldsymbol{\omega}, \boldsymbol{\omega}' \sim s} \; \frac{1}{4} \big[\cos(\boldsymbol{\omega}^T \boldsymbol{x} - \boldsymbol{\omega}'^T \boldsymbol{x}') + \cos(\boldsymbol{\omega}'^T \boldsymbol{x} - \boldsymbol{\omega}^T \boldsymbol{x}') \\
& + \; \cos(\boldsymbol{\omega}^T \boldsymbol{x} - \boldsymbol{\omega}^T \boldsymbol{x}') + \cos(\boldsymbol{\omega}'^T \boldsymbol{x} - \boldsymbol{\omega}'^T \boldsymbol{x}') \big] \\
\approx \; & \frac{1}{4D} \sum_{i=1}^{D} \Big[\cos(\boldsymbol{\omega}_i^T \boldsymbol{x} - \boldsymbol{\omega}_i'^T \boldsymbol{x}') + \cos(\boldsymbol{\omega}_i'^T \boldsymbol{x} - \boldsymbol{\omega}_i^T \boldsymbol{x}') \\
& + \; \cos(\boldsymbol{\omega}_i^T \boldsymbol{x} - \boldsymbol{\omega}_i^T \boldsymbol{x}') + \cos(\boldsymbol{\omega}_i'^T \boldsymbol{x} - \boldsymbol{\omega}_i'^T \boldsymbol{x}') \Big] \\
= \; & \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle
\end{aligned}$$

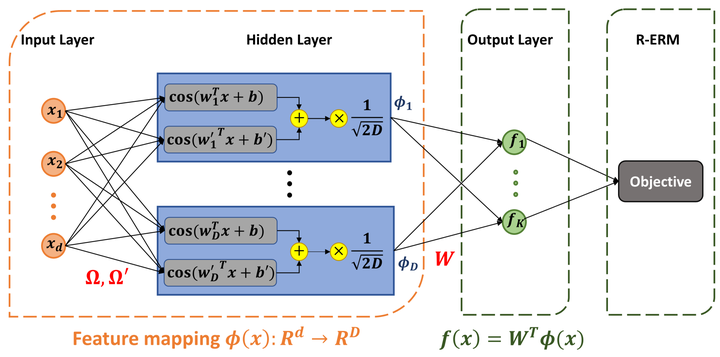
where $(\boldsymbol{\omega}_i, \boldsymbol{\omega}_i')_{i=1}^D \overset{\text{i.i.d.}}{\sim} s(\boldsymbol{\omega}, \boldsymbol{\omega}')$, the phase vectors $\boldsymbol{b}, \boldsymbol{b}'$ are drawn uniformly from $[0, 2\pi]^D$ and

$$\phi(\boldsymbol{x}) = \frac{1}{\sqrt{2D}} \left[\cos(\boldsymbol{\Omega}^T \boldsymbol{x} + \boldsymbol{b}) + \cos(\boldsymbol{\Omega}'^T \boldsymbol{x} + \boldsymbol{b}') \right].$$

The trace norm $\|\boldsymbol{W}\|_*$ and the squared Frobenius norm $\|\phi(\boldsymbol{X})\|_F^2$ exerts constraints on updating model weights $\boldsymbol{W}$ and frequency matrices $\boldsymbol{\Omega}, \boldsymbol{\Omega}'$. In this paper, we put two additional regularization terms into the ERM

$$\arg\min_{\boldsymbol{W}, \boldsymbol{\Omega}, \boldsymbol{\Omega}'} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \boldsymbol{y}_i) + \lambda_1 \|\boldsymbol{W}\|_* + \lambda_2 \|\phi(\boldsymbol{X})\|_F^2,$$

where both feature mappings $\phi(\boldsymbol{X}) \in \mathbb{R}^{D \times n}$ on all data and $f(\boldsymbol{x}_i) = \boldsymbol{W}^T \phi(\boldsymbol{x}_i) \in \mathbb{R}^D$ use the non-stationary spectral representation.



Generalization Analysis

Definition 1. *The empirical Rademacher complexity of hypothesis space $\mathcal{H}$ is*

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

where $f_k(\boldsymbol{x}_i)$ is the k-th value of the estimator $f(\boldsymbol{x}_i)$ with K outputs and ϵ_{ik}s are $n \times K$ independent Rademacher variables.

Theorem 1 (Excess Risk Bound)**.** *Assume that $B = \sup_{f \in \mathcal{H}} \|\boldsymbol{W}\|_* < \infty$ and assume the loss function ℓ is L-Lipschitz for $\mathbb{R}^K$, with probability at least $1 - \delta$, the excess risk bound holds*

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(f^*) \leq 4\sqrt{2} L \widehat{\mathcal{R}}(\mathcal{H}) + \mathcal{O}\Big(\sqrt{\frac{\log 1/\delta}{n}}\Big),$$

where $f^* \in \mathcal{H}$ is the most accurate estimator in the hypothesis space, $\widehat{f}_n$ is the empirical estimator and the empirical Rademacher complexity is

$$\begin{aligned}
\widehat{\mathcal{R}}(\mathcal{H}) \leq \; & \frac{B}{n} \sqrt{K \sum_{i=1}^{n} \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_i) \rangle} \\
= \; & \frac{B}{n} \sqrt{\frac{K}{D} \sum_{i=1}^{n} \sum_{j=1}^{D} \frac{1}{2} \Big[\cos\left((\boldsymbol{\omega}_j - \boldsymbol{\omega}_j')^T \boldsymbol{x}_i \right) + 1 \Big]}.
\end{aligned}$$

1. **The Influence of Non-Stationary Kernels.** $k(\boldsymbol{x}_i, \boldsymbol{x}_i) = \cos(\boldsymbol{\omega}^T (\boldsymbol{x}_i - \boldsymbol{x}_i)) = 1$, thus the trace of kernel matrix $\sum_{i=1}^n k(\boldsymbol{x}_i, \boldsymbol{x}_i) = n$, which corresponds to the worst cases. While for non-stationary kernels, $k(\boldsymbol{x}_i, \boldsymbol{x}_i) = \cos((\boldsymbol{\omega} - \boldsymbol{\omega}')^T \boldsymbol{x}_i) \in [-1, 1]$.

2. **Minimize the Trace Norm $\|\boldsymbol{W}\|_*$.** The convergence rate $B = \sup_{f \in \mathcal{H}} \|\boldsymbol{W}\|_* < \infty$ is also dependent on a constant B, that is the supremum of trace norm $\|\boldsymbol{W}\|_*$ in terms of the specific hypothesis space.

3. **Minimize the Squared Frobenius Norm $\|\phi(\boldsymbol{X})\|_F^2$.** Rademacher complexity is bounded by the trace of the kernel

$$\sum_{i=1}^{n} \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_i) \rangle = \sum_{i=1}^{n} \|\phi(\boldsymbol{x}_i)\|_2^2 = \|\phi(\boldsymbol{X})\|_F^2.$$

Experimental Results

		SK	NSK	SKL	NSKL	ASKL
Accuracy(↑)	segment	89.93±2.12	90.15±2.08	94.58±1.86	94.37±0.81	**95.02±1.54**
	satimage	74.54±1.35	75.15±1.38	83.61±1.08	83.74±1.34	**85.32±1.45**
	USPS	93.19±2.84	93.81±2.13	95.13±0.91	95.27±1.65	**97.76±1.14**
	pendigits	96.93±1.53	97.39±1.41	98.19±2.30	98.28±1.68	**99.06±1.26**
	letter	76.50±1.21	78.21±1.56	93.60±1.14	94.66±2.21	**95.70±1.74**
	porker	49.80±2.11	51.85±0.97	54.27±2.72	54.69±1.68	**54.85±1.28**
	shuttle	98.17±2.81	98.21±1.46	98.87±1.42	98.74±1.07	**98.98±0.94**
	MNIST	96.03±2.21	96.45±2.16	96.67±1.61	98.03±1.16	**98.26±1.78**
RMSE(↓)	abalone	10.09±0.42	9.71±0.28	8.35±0.28	**7.85±0.42**	7.88±0.16
	space_ga	11.86±0.26	11.58±0.42	11.40±0.18	11.39±0.46	**11.34±0.27**
	cpusmall	2.77±0.71	2.84±0.38	2.56±0.72	2.57±0.63	**2.42±0.48**
	cadata	50.31±0.92	51.47±0.32	47.67±0.33	47.71±0.30	**46.34±0.23**