The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Randomized Kernel Selection with Spectra of Multilevel Circulant Matrices

Lizhong Ding,[1] **Shizhong Liao,**[2] **Yong Liu,**[3] **Peng Yang,**[1] **Xin Gao**[1,*]
[1] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal, 23955-6900, Saudi Arabia
[2] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
[3] Institute of Information Engineering, CAS, Beijing, China
lizhong.ding@kaust.edu.sa, szliao@tju.edu.cn, liuyong@iie.ac.cn, {peng.yang.2, xin.gao}@kaust.edu.sa

Abstract

Kernel selection aims at choosing an appropriate kernel function for kernel-based learning algorithms to avoid either underfitting or overfitting of the resulting hypothesis. One of the main problems faced by kernel selection is the evaluation of the goodness of a kernel, which is typically difficult and computationally expensive. In this paper, we propose a randomized kernel selection approach to evaluate and select the kernel with the spectra of the specifically designed multilevel circulant matrices (MCMs), which is statistically sound and computationally efficient. Instead of constructing the kernel matrix, we construct the randomized MCM to encode the kernel function and all data points together with labels. We build a one-to-one correspondence between all candidate kernel functions and the spectra of the randomized MCMs by Fourier transform. We prove the statistical properties of the randomized MCMs and the randomized kernel selection criteria, which theoretically qualify the utility of the randomized criteria in kernel selection. With the spectra of the randomized MCMs, we derive a series of randomized criteria to conduct kernel selection, which can be computed in log-linear time and linear space complexity by fast Fourier transform (FFT). Experimental results demonstrate that our randomized kernel selection criteria are significantly more efficient than the existing classic and widely-used criteria while preserving similar predictive performance.

Introduction

Model selection of learning is the problem of choosing an appropriate hypothesis space, in which the learning algorithm searches the optimal hypothesis with the available training data, so as to avoid either underfitting or overfitting of the resulting hypothesis. For kernel-based learning algorithms, the candidate hypothesis spaces are the constrained reproducing kernel Hilbert spaces (RKHSs) (Cucker and Smale 2002), which are determined by the kernel functions and the regularization parameters. In this case, model selection is reduced to the selection of the kernel function and the regularization parameter. This paper focuses on the evaluation and selection of the kernel function, which is a fundamental and critical problem in kernel-based learning.

Kernel selection is usually to select a good kernel by minimizing or maximizing some kernel selection criteria (Bartlett, Boucheron, and Lugosi 2002; Anguita et al. 2012; Liu et al. 2017). The existing kernel selection criteria can be classified into three categories based on different inductive biases. The first category is to minimize the theoretical upper bounds of the generalization error. The upper bounds are composed of the error on data and the complexity of the hypothesis space (Bartlett, Boucheron, and Lugosi 2002). Different measures of the complexity constitute different kernel selection criteria, such as Rademacher complexity (Bartlett and Mendelson 2002), local Rademacher complexity (Cortes, Kloft, and Mohri 2013), radius-margin bound (Chapelle et al. 2002), maximum mean discrepancy (MMD) (Sriperumbudur et al. 2009; Gretton et al. 2012a; 2012b; Song et al. 2012), effective dimensionality (Zhang 2005; Bach 2013), eigenvalues ratio (Liu and Liao 2015), and the covering number (Ding and Liao 2014b). The second category is to maximize the similarity between the kernel matrix and the label matrix. The criteria in this category are two-stage kernel selection criteria (Cortes, Mohri, and Rostamizadeh 2010), which do not require the training of the learning algorithms in the kernel selection step. The representative criterion is the kernel target alignment (KTA) (Cristianini et al. 2002). Feature space-based measure (FSM) (Nguyen and Ho 2007) and centered KTA (CKTA) (Cortes, Mohri, and Rostamizadeh 2010) were proposed to improve the performance of KTA. The third category is to minimize the statistical experimental errors, including hold-out method, cross validation (CV), leave-one-out (LOO) and Bootstrap. The CV error is probably the most commonly used criterion in the machine learning community and the LOO error gives an almost unbiased estimate of the generalization error (Luntz and Brailovsky 1969). However, CV and LOO require training the learning algorithm for every candidate parameter many times, unavoidably bringing high computational burdens. For the sake of efficiency, some approximate CV approaches were proposed, such as generalized cross validation (GCV) (Golub, Heath, and Wahba 1979), efficient LOO (Cawley and Talbot 2010), and Bouligand influence function CV (BIFCV) (Liu, Jiang, and Liao 2014).

The computational complexities of the existing kernel selection criteria are at least quadratic in the number of examples l, i.e., $O(l^2)$, and usually $O(l^3)$. This kind of scalability

*All correspondence should be addressed to Xin Gao.

is prohibitive for the large-scale supervised learning. It is worth noting that linear-time criteria via MMD were proposed in (Gretton et al. 2012a; 2012b), which, however, are for two-sample tests but not for the general supervised learning. For the criterion in (Gretton et al. 2012a), one needs to calculate the kernel functions between two different sets of samples and for the criterion in (Gretton et al. 2012b), it requires the estimation of the asymptotic probability of a Type II error. On the other hand, a kernel selection criterion is not required to be an unbiased estimate of the generalization error, and instead the primary requirement is merely for the minimum of the kernel selection criterion to provide a reliable indication of the minimum of the generalization error in the kernel parameter space. Therefore, we argue that it is sufficient to calculate randomized criteria with theoretical guarantees that can discriminate the optimal kernel from other candidates. The above two considerations drive the study of this paper.

In this paper, we propose a randomized kernel selection approach to evaluate and select the kernel in the space of the spectra of the specifically designed multilevel circulant matrices (MCMs). In contrast to the traditional kernel selection methods, we do not use the kernel function and the data to generate a kernel matrix. Instead, for each candidate kernel we explicitly construct a MCM encoding the kernel function and all data points together with their labels. The reason why we choose MCM is that the built-in periodicity of MCM allows the multidimensional fast Fourier transform (mFFT) to be utilized in calculating its eigenvalues and eigenvectors in quasi-linear time complexity, which is much faster than the eigen-decomposition of the kernel matrix. More specifically, we build a one-to-one correspondence between the candidate kernel functions and the spectra of the randomized MCMs by using Fourier transform (FT) twice; the first FT is from the harmonic analysis of random features (Rahimi and Recht 2008; 2009) and the second one is from the built-in periodicity of the MCMs. In the space of spectra of MCMs, we define a series of randomized criteria based on different inductive biases to conduct approximate kernel selection in a log-linear time and linear space complexity. Theoretically, we analyze the statistical properties of the randomized MCMs and the approximate kernel selection criteria to qualify their utility in kernel selection. Empirically, we provide the experimental evidence that our criteria are significantly more efficient than the existing ones while preserving similar predictive performance.

Related work

Circulant matrix (CM) has been adopted in random projection and kernel matrix approximation. On the basis of the Johnson-Lindenstrauss lemmas (Hinrichs and Vybíral 2011; Vybíral 2011), circulant random projection has been successfully used in binary embedding (Yu et al. 2014) and parameter redundancy of deep networks (Cheng et al. 2015). For kernel matrix approximation, an ingenious algorithm (Song and Xu 2010) was proposed to construct MCMs as the approximations of kernel matrices. However, the constructions of these existing CMs or MCMs are all independent of the data. For circulant random projection, the CM is completely random and for MCM in (Song and Xu 2010; Ding and Liao 2014a), only the kernel function is used. In this paper, we construct a novel type of data-dependent MCMs, which are different from the existing ones and fit the needs of kernel selection.

Random features were proposed to approximate non-linear kernel functions. The seminal work, random Fourier features (Rahimi and Recht 2008), focuses on approximating shift-invariant kernels. Several approaches were also proposed to approximate other types of kernels, such as additive kernels (Vedaldi and Zisserman 2012) and dot product kernels (Kar and Karnick 2012). Recently, attention has been paid on improving the approximation quality of random features (Hamid and Xiao 2014; Yang et al. 2014) and accelerating the approximation procedure (Le, Sarlòs, and Smola 2013). In the fist step of constructing our randomized MCMs, we adopt the harmonic analysis of random features (Rahimi and Recht 2008; Kar and Karnick 2012). However, existing techniques for random features assume a user-defined kernel as the input and leave the kernel selection problem to the user. Selecting a good kernel is a more challenging problem than approximating a known kernel. This paper focuses on the evaluation and selection of the kernel function.

Notations and Preliminaries

In this section, we introduce the adopted notations and the notion of multilevel circulant matrices.

We consider a continuous, symmetric and positive kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Cucker and Smale 2002). That is, for any finite set $\{x_0, \ldots, x_{l-1}\} \subseteq \mathcal{X}$, the matrix $\mathbf{K} = [\kappa(x_i, x_j)]_{i,j=0}^{l-1}$ is symmetric and positive definite (SPD). The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\kappa$ associated with the kernel κ is defined as $\mathcal{H}_\kappa = \overline{\text{span}}\{\kappa(x, \cdot) : x \in \mathcal{X}\}$. $\mathbf{AB}$, $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{A} \circ \mathbf{B}$ denote matrix multiplication, Kronecker product and Hadamard product between $\mathbf{A} \in \mathbb{R}^{l \times l}$ and $\mathbf{B} \in \mathbb{R}^{l \times l}$, respectively.

In the following, we introduce the notion of multilevel circulant matrices. For any positive integer m, let $[m] = \{0, 1, \ldots, m-1\}$. For a fixed positive integer p, and $\boldsymbol{m} = (m_0, m_1, \ldots, m_{p-1}) \in \mathbb{N}^p$, we write the continued product as $\Pi_{\boldsymbol{m}} = m_0 m_1 \ldots m_{p-1}$ and the Cartesian product as $[\boldsymbol{m}] = [m_0] \times [m_1] \times \cdots \times [m_{p-1}]$. A 1-level circulant matrix (CM) $\mathbf{C}$ is a matrix having the form

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{m-1} & \ldots & c_1 \\ c_1 & c_0 & \ldots & c_2 \\ \vdots & \vdots & \ddots & \vdots \\ c_{m-1} & c_{m-2} & \ldots & c_0 \end{bmatrix},$$

where each column is a cyclic shift of its left column. The (i, j)-th entry of $\mathbf{C}$ satisfies $C_{i,j} = c_{i-j(\text{mod } m)}$. Since $\mathbf{C}$ is fully determined by its first column, we write

$$\mathbf{C} = \text{circ}[c_i : i \in [m]].$$

A multilevel circulant matrix (MCM) is defined recursively. For any positive integer p, a $(p+1)$-level CM is a block CM whose blocks are p-level CMs. For $\boldsymbol{m} \in \mathbb{N}^p$, we use multi-indices $\boldsymbol{i} = (i_0, \ldots, i_{p-1}), \boldsymbol{j} = (j_0, \ldots, j_{p-1}) \in [\boldsymbol{m}]$

to locate the entries of a p-level CM $\mathbf{A}_{\boldsymbol{m}}$. According to (Tyrtyshnikov 1996), for $\boldsymbol{m} \in \mathbb{N}^p$, $\mathbf{A}_{\boldsymbol{m}} = [A_{\boldsymbol{i},\boldsymbol{j}} : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$ is a p-level CM if, for any $\boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]$,

$$A_{\boldsymbol{i},\boldsymbol{j}} = a_{i_0-j_0 (\mathrm{mod}\ m_0),\ldots,i_{p-1}-j_{p-1} (\mathrm{mod}\ m_{p-1})}.$$

$\mathbf{A}_{\boldsymbol{m}}$ is completely determined by its first column. We write $\mathbf{A}_{\boldsymbol{m}} = \mathrm{circ}_{\boldsymbol{m}}[a_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$.

Spectrum Space on Randomized MCMs for Kernel Selection: Definitions and Theories

In this section, we define a kind of randomized MCMs encoding the information of the kernel function and all data points together with their labels, which can be considered as a higher-dimensional approximation of the labeled kernel matrix $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \circ \mathbf{K}$ with excellent theoretical and computational virtues. Since all MCMs of the same dimension have the same eigenvectors (Gray 1972), we can map the defined MCMs to their spectra by Fourier transform. We build a one-to-one correspondence between all candidate kernel functions and the spectra of the randomized MCMs. We will approximately select the optimal kernel in the space of the spectra of the randomized MCMs. We theoretically justify the rationality of the randomized MCMs and the spectrum space for kernel selection.

We first consider $\mathcal{X} = \mathbb{R}^d$ and shift-invariant kernels $\kappa(x, y) = \zeta(x - y)$, where ζ is an integrable function from $\mathbb{R}^d$ to $\mathbb{R}$. We know that shift-invariant kernels are positive definite if and only if the Fourier transform of ζ is always a non-negative real number (Bochner 1933). If $\hat{\zeta}(w) = \int_{\mathbb{R}^d} \zeta(x) e^{-iw^{\mathrm{T}}x} dx \in \mathbb{R}_+$, then

$$k(x, y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{\zeta}(w) e^{iw^{\mathrm{T}}(x-y)} dw.$$

Inspired by (Rahimi and Recht 2008; 2009), by sampling $w_0, \ldots, w_{D-1}$ from a density proportional to $\hat{\zeta}(w) \in \mathbb{R}_+$ and $b_0, \ldots, b_{D-1}$ uniformly in $[0, 2\pi]$, we can define a CM $\mathbf{C}(x_i) \in \mathbb{R}^{D\times D}$ for x_i, $i \in [l]$,

$$\mathbf{C}(x_i) = \frac{\sqrt{2}}{D} \mathrm{circ} \begin{pmatrix} \cos(x_i^{\mathrm{T}} w_0 + b_0) \\ \cos(x_i^{\mathrm{T}} w_1 + b_1) \\ \vdots \\ \cos(x_i^{\mathrm{T}} w_{D-1} + b_{D-1}) \end{pmatrix}.$$

The rationality of the definition of $\mathbf{C}(x_i)$ will be shown in the end of this section. We can also define CMs for dot product kernels based on Schoenberg theorem (Schoenberg 1942; Kar and Karnick 2012).

For $i \in [l]$, we further define a labeled CM as $\dot{\mathbf{C}}(x_i) = y_i \mathbf{C}(x_i)$, where $y_i \in \{-1, 1\}$ for classification or $y_i \in \mathbb{R}$ for regression. In order to involve all data information, we cycle all $\dot{\mathbf{C}}(x_i)$ for x_i to define an MCM $\mathbf{U}_{\mathbf{m}}$ of order $\boldsymbol{m} = (l, D)$,

$$\mathbf{U}_{\mathbf{m}} = \begin{pmatrix} \dot{\mathbf{C}}(x_0) & \dot{\mathbf{C}}(x_{l-1}) & \ldots & \dot{\mathbf{C}}(x_1) \\ \dot{\mathbf{C}}(x_1) & \dot{\mathbf{C}}(x_0) & \ldots & \dot{\mathbf{C}}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\mathbf{C}}(x_{l-1}) & \dot{\mathbf{C}}(x_{l-2}) & \ldots & \dot{\mathbf{C}}(x_0) \end{pmatrix}.$$

$\mathbf{U}_{\mathbf{m}}$ is actually a $\Pi_{\boldsymbol{m}} \times \Pi_{\boldsymbol{m}}$ matrix encoding the kernel function and all data points together with their labels.

According to (Tyrtyshnikov 1996), we know that $\mathbf{U}_{\boldsymbol{m}}$ is an MCM of order $\boldsymbol{m}$ if and only if

$$\mathbf{U}_{\boldsymbol{m}} = \frac{1}{\Pi_{\boldsymbol{m}}} \mathbf{\Phi}^{\mathrm{H}} \mathrm{diag}(\mathbf{\Phi}\boldsymbol{u}) \mathbf{\Phi}, \tag{1}$$

where H denotes the conjugate transpose, $\mathrm{diag}(w)$ denotes the diagonal matrix with w on the main diagonal, and $\mathbf{\Phi} = \mathbf{F}_l \otimes \mathbf{F}_D$, with $\mathbf{F}_m = \left[\mathrm{e}^{\frac{2\pi \mathrm{i}}{m} st} : s, t \in [m]\right]$ for $m \in \mathbb{N}$. The eigenvalues of $\mathbf{U}_{\boldsymbol{m}} = \mathrm{circ}_{\boldsymbol{m}}[u_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$ are given by

$$v_{\boldsymbol{j}} = \sum_{\boldsymbol{i} \in [\boldsymbol{m}]} u_{\boldsymbol{i}} \mathrm{e}^{2\pi \mathrm{i} \sum_{s\in[p]} \frac{i_s j_s}{m_s}}, \quad \boldsymbol{j} \in [\boldsymbol{m}]. \tag{2}$$

For a set of candidate kernels $\mathcal{K} = \{\kappa^{(i)} : i \in [N]\}$, we can construct an MCM $\mathbf{U}_{\boldsymbol{m}}^{(i)}$ for each kernel $\kappa^{(i)}$, and then we have a candidate set $\mathcal{U} = \{\mathbf{U}_{\boldsymbol{m}}^{(i)} : i \in [N]\}$ for kernel selection. For each $\mathbf{U}_{\boldsymbol{m}}^{(i)}$, we can obtain the spectrum $\boldsymbol{v}^{(i)} = [v_{\boldsymbol{j}}^{(i)}, \boldsymbol{j} \in [\boldsymbol{m}]]$ according to (2). Since the eigenvectors of an MCM are the Kronecker product of Fourier matrices, all $\mathbf{U}_{\boldsymbol{m}}^{(i)}$ for $i \in [N]$ have the same eigenvectors (Gray 1972). The discriminative information of different kernels are encoded into the spectrum $\boldsymbol{v}^{(i)}$. For $\mathcal{V} = \{\boldsymbol{v}^{(i)} : i \in [N]\}$, we will (approximately) select a good kernel as

$$\kappa^* \approx \arg\max_{i\in[N]} \mathcal{C}(\boldsymbol{v}^{(i)}) \quad \text{or} \quad \kappa^* \approx \arg\min_{i\in[N]} \mathcal{C}(\boldsymbol{v}^{(i)}), \tag{3}$$

where the specific forms of the kernel selection criterion $\mathcal{C}(\boldsymbol{v}^{(i)})$ will be given in the next section.

Remark We can easily extend our method from selecting a kernel to optimizing the combination weights of base kernels for multiple kernel learning. The optimization strategies with MCMs have been given in (Ding and Liao 2017) for the L_1 and L_2 regularization (Cortes, Mohri, and Rostamizadeh 2009) of the combination weights.

The following theoretical guarantees show the rationality of the MCMs and $\mathcal{V} = \{\boldsymbol{v}^{(i)} : i \in [N]\}$ for kernel selection. We define the inner product between $\mathbf{C}(x_i)$ and $\mathbf{C}(x_j)$ as $\langle \mathbf{C}(x_i), \mathbf{C}(x_j) \rangle = \mathbf{1}_D^{\mathrm{T}} (\mathbf{C}(x_i) \circ \mathbf{C}(x_j)) \mathbf{1}_D$, where $\mathbf{1}_D$ denotes the vector of all ones of length D. Lemma 1 reveals the relationship between the matrix multiplication and the Hadamard product of two CMs.

Lemma 1. *For any $i, j \in [l]$,*

$$\mathbb{E}\left[\mathbf{1}_D^{\mathrm{T}} (\mathbf{C}(x_i)\mathbf{C}(x_j)) \mathbf{1}_D\right] = \mathbb{E}\left[\langle \mathbf{C}(x_i), \mathbf{C}(x_j) \rangle\right].$$

Based on Lemma 1, we can prove Theorem 1, which reveals that $\langle \mathbf{C}(x_i), \mathbf{C}(x_j) \rangle$ is equivalent to $\kappa(x_i, x_j)$ in expectation.

Theorem 1. *For any $i, j \in [l]$,*

$$\mathbb{E}\left[\langle \mathbf{C}(x_i), \mathbf{C}(x_j) \rangle\right] = \kappa(x_i, x_j).$$

Theorem 2 shows the concentration bound between $\langle \mathbf{C}(x_i), \mathbf{C}(x_j) \rangle$ and $\kappa(x_i, x_j)$, which reveals convergence speed between these two values. This is the result for the Gaussian kernel $\kappa(x - x') = \exp(-\frac{\|x-x'\|_2^2}{2\sigma^2})$.

Theorem 2. *For any $i, j \in [l]$, denoting $\Delta = \|x_i - x_j\|_2^2$, we have*

$$\Pr\left\{ |\langle \mathbf{C}(x_i), \mathbf{C}(x_j)\rangle - \kappa(x_i, x_j)| \geq \varepsilon \right\} \leq \frac{\mathrm{Var}}{D^2\varepsilon},$$

where

$$\mathrm{Var} = \frac{1}{D} + \frac{1}{2D}\exp\left(-\frac{2\Delta}{\sigma^2}\right) + \frac{D-2}{D}\exp^2\left(-\frac{\Delta}{\sigma^2}\right).$$

We represent the eigen-decomposition of $\mathbf{K}$ as $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{\Sigma}$ and $\mathbf{U}$ denote the eigenvalues and eigenvectors of $\mathbf{K}$, respectively. Theorem 3 establishes the statistical relation between the eigenvalues of the MCM $\mathbf{U_m}$ and the eigenvalues of the kernel matrix $\mathbf{K}$. The reasons why we derive Theorem 3 are: 1) the spectrum of the kernel plays an important role in evaluating the goodness of the kernel; 2) $\boldsymbol{y}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}\boldsymbol{y}$ is a critical component for various kernel selection criteria, which will be shown in the next section. Theorem 3 shows that $l^{-1}\mathbf{1}_{\Pi_{\boldsymbol{m}}}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{H}}\mathrm{diag}(v_{\boldsymbol{j}}^2)\mathbf{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}$ is an unbiased estimator of $\boldsymbol{y}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}\boldsymbol{y}$.

Theorem 3. *For $\boldsymbol{j} \in [\boldsymbol{m}]$,*

$$\mathbb{E}\left[l^{-1}\mathbf{1}_{\Pi_{\boldsymbol{m}}}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{H}}\mathrm{diag}(v_{\boldsymbol{j}}^2)\mathbf{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}\right] = \boldsymbol{y}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}\boldsymbol{y}.$$

Remark For random Fourier features, we denote $z_i := [\cos(x_i^{\mathrm{T}}w_0 + b_0), \ldots, \cos(x_i^{\mathrm{T}}w_{D-1} + b_{D-1})] \in \mathbb{R}^D$ and $\mathbf{Z} := [z_0 \ldots z_l] \in \mathbb{R}^{D\times l}$. Here we discuss the difference between $\mathbf{Z}$ and $\mathbf{U_m}$. For the MCM $\mathbf{U_m}$, we can explicitly represent the spectrum of $\mathbf{U_m}$ using its first column, as shown in (2). If we use $\mathbf{Z}$, we should conduct SVD on $\mathbf{Z}$ to obtain the eigenvalues. But the eigenvalues are not in an explicit form. This difference is similar to the difference between CUR decomposition and SVD (please see (Mahoney and Drineas 2009) for details). As shown in Theorem 2, the concentration bound converges with the rate $O(D^{-2})$, which is faster than the direct use of $\mathbf{Z}$. If we use $\mathbf{Z}$, to guarantee the same convergence rate, we need to sample D^2 random vectors from a density proportional to $\hat{t}(w) \in \mathbb{R}_+$. The complexity of FFT for solving the eigen-decomposition of $\mathbf{U_m}$ is $O(\Pi_{\boldsymbol{m}}\ln(\Pi_{\boldsymbol{m}}))$, that is $O(lD\ln(lD))$. The complexity of SVD on $\mathbf{Z} \in D^2 \times l$ is lD^4. When $l < \frac{3e^D}{D}$, $O(lD\ln(lD)) < lD^4$. In practice, we can also adopt the direct stack $\mathbf{Z}$ to conduct randomized kernel selection, but sometimes it may produce poor results for the same value of D, which will be shown in experiments.

Randomized Kernel Selection Criteria on the Spectra of the MCMs

The spectrum of the kernel plays an important role in evaluating the goodness of the kernel for all the three categories of kernel selection criteria shown in Introduction. With the theoretical support provided by Theorem 1, 2 and 3, we design a series of randomized kernel selection criteria for the three categories in this section. These criteria alleviate the computational bottleneck faced by existing kernel selection approaches and provide a solution with log-linear time and linear space complexity.

We start from the first category. In the upper bounds of the generalization error that are composed of the error on data and the complexity of the hypothesis space, the complexity term for kernel-based learning can usually be represented in the spectrum of the kernel matrix, such as effective dimensionality (Zhang 2005; Bach 2013) and local Rademacher complexity (Cortes, Kloft, and Mohri 2013). Here we adopt effective dimensionality as a case to design the randomized kernel selection criterion, which can also be extended to other measures (Cortes, Kloft, and Mohri 2013).

We observe corrupted response $y_i = \dot{y}_i + \xi_i$, $1 \leq i \leq l$, where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_i]^{\mathrm{T}}$ are random variables with mean 0 and finite covariance matrix $\mathbf{C}$, and $\dot{\boldsymbol{y}} = [\dot{y}_1, \ldots \dot{y}_l]^{\mathrm{T}}$ is the underlying true output. We consider the regularized empirical error $\mathcal{E}(f) = \frac{1}{l}\sum_{i=1}^{l}(f(x_i) - y_i)^2 + \mu\|f\|_{\mathcal{H}_\kappa}^2$, where μ is the regularization parameter. The optimal function $f_\kappa = \arg\min_{f\in\mathcal{H}_\kappa}\mathcal{E}(f)$. By the representer theorem (Kimeldorf and Wahba 1970), we have $f_\kappa = \sum_{i=1}^{l}\alpha_i\kappa(x_i, \cdot)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^{\mathrm{T}} = (\mathbf{K} + \mu l\mathbf{I})^{-1}\boldsymbol{y}$. Therefore, $\|f_\kappa\|_{\mathcal{H}_\kappa}^2 = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} = \boldsymbol{y}^{\mathrm{T}}\mathbf{K}_\mu^{-1}\mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$, where $\mathbf{K}_\mu = \mathbf{K} + \mu l\mathbf{I}$. Denoting $\boldsymbol{f}_\kappa = (f_\kappa(x_1), \ldots, f_\kappa(x_l))^{\mathrm{T}}$, we have $\boldsymbol{f}_\kappa = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}_\mu^{-1}\boldsymbol{y}$, which is referred to as an estimate of $\dot{\boldsymbol{y}}$. The expected error of $\boldsymbol{f}_\kappa$ (Bach 2013) is

$$\begin{aligned}\frac{1}{l}\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{f}_\kappa - \dot{\boldsymbol{y}}\|^2 &= \frac{1}{l}\|\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{f}_\kappa - \dot{\boldsymbol{y}}\|^2 + \frac{1}{l}\mathrm{trace}(\mathrm{var}_{\boldsymbol{\xi}}(\boldsymbol{f}_\kappa)) \\ &= \underbrace{\mu^2 l\dot{\boldsymbol{y}}^{\mathrm{T}}\mathbf{K}_\mu^{-2}\dot{\boldsymbol{y}}}_{\mathrm{bias}(\mathbf{K})} + \underbrace{\frac{1}{l}\mathrm{trace}(\mathbf{C}\mathbf{K}^2\mathbf{K}_\mu^{-2})}_{\mathrm{variance}(\mathbf{K})}.\end{aligned}$$

The bias term is the error on data. The variance term controls the model complexity, which is the effective dimension $\mathrm{trace}(\mathbf{C}\mathbf{K}^2\mathbf{K}_\mu^{-2}) \leq \mathrm{trace}(\mathbf{C}\mathbf{K}\mathbf{K}_\mu^{-1})$ (Zhang 2005). By minimizing the sum of bias and variance, we can minimize the generlization error (Zhang 2005). For $\mathbf{C} = \sigma^2\mathbf{I}$, we can define a kernel selection criterion as $\mathcal{C}_1(\mathbf{K}) = \mu^2 l\dot{\boldsymbol{y}}^{\mathrm{T}}\mathbf{K}_\mu^{-2}\dot{\boldsymbol{y}} + \frac{\sigma^2}{l}\mathrm{trace}(\mathbf{K}\mathbf{K}_\mu^{-1})$, because $\mathcal{C}_1(\mathbf{K})$ only depends on the kernel κ for a fixed μ and we can select the optimal kernel with the minimum generlization error.

However, the computational complexity of $\mathcal{C}_1(\mathbf{K})$ is $O(l^3)$. With the theoretical guarantee of the last section, we define a randomized kernel selection criterion,

$$\begin{aligned}\mathcal{C}_1(\boldsymbol{v}) = &\frac{1}{\Pi_{\boldsymbol{m}}}\mathbf{1}_{\Pi_{\boldsymbol{m}}}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{H}}\mathrm{diag}\left(\frac{1}{(v_{\boldsymbol{j}}^2 + \mu l)^2}, \boldsymbol{j} \in [\boldsymbol{m}]\right)\mathbf{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}} \\ &+ \sum_{\boldsymbol{j}\in[\boldsymbol{m}]}\frac{v_{\boldsymbol{j}}^2}{v_{\boldsymbol{j}}^2 + \mu l},\end{aligned} \tag{4}$$

where the spectrum $\boldsymbol{v} = [v_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]]$. According to (1), $\mathbf{\Phi}$ is the Kronecker product of Fourier matrices, which means that for any vector $x = [x_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$, $\mathbf{\Phi}x$ is the multidimensional discrete Fourier transform (mDFT) of x. That is, $\mathbf{\Phi}x$ can be calculated by multidimensional fast Fourier transform (mFFT) (Singleton 1969). According to (2), $\boldsymbol{v} = [v_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]]$ can be calculated using the mFFT of $[u_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$. $\mathbf{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}$ and $\boldsymbol{v}$ can both be calculated using mFFT. Hence, the time complexity of $\mathcal{C}_1(\boldsymbol{v})$ is

$O(\Pi_{\boldsymbol{m}} \ln(\Pi_{\boldsymbol{m}}))$. Since only the first column of $\mathbf{U}_{\boldsymbol{m}}$ needs to be stored, the space complexity is $O(\Pi_{\boldsymbol{m}})$. In the second term of (4), we use the approximation of labeled kernel matrix $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \circ \mathbf{K}$ to measure the effective dimension, which is not exactly the same as in $\mathcal{C}_1(\mathbf{K})$ and involves more data information for kernel selection.

Now we consider the second category of kernel selection criteria, the similarity-based criteria. The most representative one is the kernel target alignment (KTA) (Cristianini et al. 2002), $\mathrm{KTA}(\mathbf{K}) = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{K}\boldsymbol{y}}{l\|\mathbf{K}\|_{\mathrm{F}}}$. We define a randomized kernel selection criterion for KTA as

$$\mathcal{C}_2(\boldsymbol{v}) = \frac{\left(\boldsymbol{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}\right)^{\mathrm{H}} \mathrm{diag}\left(v_{\boldsymbol{j}}^2, \boldsymbol{j} \in [\boldsymbol{m}]\right) \boldsymbol{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}}{\Pi_{\boldsymbol{m}}^3 \sum_{\boldsymbol{j}\in[\boldsymbol{m}]} v_{\boldsymbol{j}}^2}. \quad (5)$$

We can extend (5) to MMD for binary classification (Sriperumbudur et al. 2009; Song et al. 2012). Let $l_+(l_-)$ denote the number of the positive (negative) data points. For $i \in [l]$, we write $\bar{y}_i = 1/l_+$, if $y_i = +1$ and $\bar{y}_i = -1/l_-$ otherwise. We can define an MCM $\bar{\mathbf{U}}_{\boldsymbol{m}}$, where $\bar{\mathbf{C}}(x_i) = \bar{y}_i \mathbf{C}(x_i)$ for $i \in [l]$. The randomized criterion is $\mathcal{C}(\bar{\boldsymbol{v}}) = \frac{1}{\Pi_{\boldsymbol{m}}^2}\left(\boldsymbol{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}\right)^{\mathrm{H}} \mathrm{diag}\left(\bar{\boldsymbol{v}}^2\right) \boldsymbol{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}$, where $\bar{\boldsymbol{v}}$ is the spectrum of $\bar{\mathbf{U}}_{\boldsymbol{m}}$.

We consider the third category of kernel selection criteria, the experimental criteria. We adopt the most commonly used criteria, the CV error and the LOO error, to discuss the definition of randomized criteria. In (Cawley and Talbot 2010), an efficient closed-form of LOO (ELOO) was proposed, which computes the LOO error in $O(l^3)$ time complexity for kernel-based learning instead of $O(l^4)$ of the direct LOO procedure. The most time consuming step in ELOO is to solve the inverse of $\mathbf{P} = [\mathbf{K}_\mu, \mathbf{1}_l; \mathbf{1}_l^{\mathrm{T}}, 0]$. Using the block matrix inversion formula, $\mathbf{P}^{-1} = [\mathbf{K}_\mu^{-1} + c^{-1}\mathbf{K}_\mu^{-1}\mathbf{1}_l\mathbf{1}_l^{\mathrm{T}}\mathbf{K}_\mu^{-1}, -c^{-1}\mathbf{K}_\mu^{-1}\mathbf{1}_l; -c^{-1}\mathbf{1}_l^{\mathrm{T}}\mathbf{K}_\mu^{-1}, c^{-1}]$, where $c = -\mathbf{1}_l^{\mathrm{T}}\mathbf{K}_\mu^{-1}\mathbf{1}_l$. We can efficiently solve $\mathbf{K}_\mu^{-1}\mathbf{1}_l$ with an unlabeled version of $\mathbf{U}_{\mathbf{m}}$, denoted as $\hat{\mathbf{U}}_{\boldsymbol{m}}$,

$$\begin{aligned}&\left(\hat{\mathbf{U}}_{\boldsymbol{m}}\hat{\mathbf{U}}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}\right)^{-1} \mathbf{1}_{\Pi_{\boldsymbol{m}}} \\ &= \frac{1}{\Pi_{\boldsymbol{m}}}\boldsymbol{\Phi}^{\mathrm{H}}\mathrm{diag}\left(\frac{1}{\hat{v}_{\boldsymbol{j}}^2 + \mu l} : \boldsymbol{j} \in [\boldsymbol{m}]\right)\boldsymbol{\Phi}\mathbf{1}_{\Pi_{\boldsymbol{m}}}.\end{aligned} \quad (6)$$

We denote $\mathcal{C}_3(\hat{\boldsymbol{v}})$ as the approximate ELOO error. For k-fold CV, BIFCV (Liu, Jiang, and Liao 2014) computes the CV error in $O(l^3 + kl^2)$ time complexity instead of $O(kl^3)$ of the direct k-fold CV procedure. In BIFCV, we need to solve $\mathbf{K}_\mu^{-1}\boldsymbol{\eta}$ and $\mathbf{K}\boldsymbol{\theta}$ (please see (Liu, Jiang, and Liao 2014) for the detailed forms of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$), which can both be randomly computed in log-linear complexity following (6).

Experiments

We first verify the effectiveness of the randomized kernel selection criteria. Effectiveness includes efficiency and generalization, where the former is measured by the average computational time for kernel selection and the latter is measured by the classification accuracy of the learned hypothesis with the selected kernel over the test set.

We conduct experiments to compare the effectiveness of the randomized kernel selection criteria $\mathcal{C}_1(\boldsymbol{v})$, $\mathcal{C}_2(\boldsymbol{v})$ and $\mathcal{C}_3(\hat{\boldsymbol{v}})$ with different kinds of classic and widely-used baselines, including kernel target alignment (KTA) (Cristianini et al. 2002), feature space-based measure (FSM) (Nguyen and Ho 2007), centered kernel target alignment (CKTA) (Cortes, Mohri, and Rostamizadeh 2010), maximum mean discrepancy for binary classification[1] (MMD-B) (Sriperumbudur et al. 2009; Song et al. 2012), effective dimension $\mathcal{C}_1(\mathbf{K})$ (Zhang 2005), 5-fold cross validation (CV), efficient leave-one-out (ELOO) (Cawley and Talbot 2010) and ELOO with Bayesian regularisation (ELOO-BR) (Cawley and Talbot 2007). The complexities of KTA, CKTA, FSM and MMD-B are $O(l^2)$ and complexities of $\mathcal{C}_1(\mathbf{K})$, 5-fold CV, ELOO and ELOO-BR are $O(l^3)$.

The set of Gaussian kernels $\kappa(x - x') = \exp(-\gamma\|x - x'\|_2^2)$ with a variable bandwidth parameter $\gamma \in \{2^i, i = -8, -7, \ldots, 5, 6\}$ is adopted as the candidate kernel set. The parameter D in the randomized criteria is an important parameter both for the approximation quality and computational efficiency ($O(\Pi_{\boldsymbol{m}} \ln(\Pi_{\boldsymbol{m}})) = O(lD\ln(lD))$). We conduct experiments for different values of D (50, 100, 200, 500, 1000, 2000). Finally, we fix $D = 100$, because it is enough to demonstrate the effectiveness of randomized kernel selection criteria[2]. We adopt least square support vector machine (LSSVM) as the learning algorithm. Since the focus of this work is not on tuning the regularization parameter, it is set as a fixed value 1. A variety of datasets that cover the number of data points ranging from 7200 up to more than 245000 and the number of features ranging from 3 up to 47,236 are selected from the UCI dataset repository[3], the LIBSVM dataset repository[4], and the KEEL dataset repository[5]. We randomly partition each dataset into two parts, with 50% of the data randomly chosen for training and the rest reserved for testing. We conduct kernel selection by minimizing or maximizing the kernel selection criteria on the training set. After the kernel selection step, LSSVM is trained with the obtained optimal kernel still on the training set. Finally, the performance of the trained model is measured on the test set. We repeat each experiment 30 times to estimate the statistical significance of differences in performance using the z statistic (Cawley and Talbot 2007), where $z = 1.64$ corresponds to a 95% significance level (Cawley and Talbot 2007). The results are shown in Table 1, in which "NA", not available, means that none of the 30 runs can be completed in 120 hours; "OM", out of memory, means that the experiment needs more than 1TB memory, which cannot be run successfully on the largest memory node of our computing cluster. We can see that our kernel selection criteria are significantly

[1]Note that the linear approximations of MMD in (Gretton et al. 2012a; 2012b) were for the purpose of testing if two sets of samples are generated from the same distribution, which needs to calculate the kernel function between two sets of samples. The MMD version for binary classification we chose here is from (Sriperumbudur et al. 2009; Song et al. 2012).

[2]The results for the values of D bigger than 100 are similar.

[3]http://www.ics.uci.edu/∼mlearn/MLRepository.html

[4]http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[5]http://sci2s.ugr.es/keel/datasets.php

Table 1: Comparison of the average accuracy and time (seconds) over the 30 runs of different criteria.

datasets (♯ examples, ♯ features)	$\mathcal{C}_1(\boldsymbol{v})$		$\mathcal{C}_2(\boldsymbol{v})$		$\mathcal{C}_3(\hat{\boldsymbol{v}})$		KTA	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
thyroid (7200, 21)	98.26%	8.65	98.29%	9.23	98.24%	13.24	97.84%	76.22
mushrooms (8124, 112)	99.92%	44.48	99.90%	45.21	99.82%	78.67	100.00%	278.19
coil2000 (9822, 85)	93.84%	119.46	93.86%	116.18	93.87%	221.40	94.10%	814.96
penbased (10992, 16)	99.81%	27.79	99.01%	25.83	99.69%	44.61	99.74%	158.79
rcv (20242, 47236)	95.37%	3695.63	95.59%	3527.24	95.39%	3601.41	NA	NA
adult (32561, 123)	84.17%	61.02	84.29%	64.21	84.42%	91.58	84.72%	4287.95
w8a (49749, 300)	98.11%	3865.91	98.24%	3890.73	97.84%	4718.15	97.80%	21715.09
cod-rna (59535, 8)	94.75%	93.23	94.88%	96.43	94.95%	150.18	94.82%	3559.34
fars (100968, 29)	85.46%	273.24	85.75%	287.59	86.01%	549.69	OM	OM
skin (245057, 3)	99.24%	4527.78	99.11%	4619.76	98.87%	8661.18	OM	OM

datasets	FSM		CKTA		MMD-B		$\mathcal{C}_1(\mathbf{K})$	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
thyroid	98.62%	1412.09	98.41%	1219.69	98.64%	61.72	98.53%	5994.80
mushrooms	93.21%	1987.76	99.65%	1621.89	99.97%	251.00	99.98%	8819.04
coil2000	93.72%	2961.75	94.02%	2305.35	94.01%	672.49	94.04%	15367.52
penbased	99.75%	3310.80	99.72%	2656.29	99.66%	113.13	99.76%	21144.28
rcv	NA	NA	NA	NA	NA	NA	NA	NA
adult	82.89%	27854.28	84.32%	24609.40	84.91%	3930.97	NA	NA
w8a	98.14%	61377.31	97.75%	52636.91	98.55%	21110.3	NA	NA
cod-rna	95.14%	58111.03	95.32%	70559.17	95.27%	2336.34	NA	NA
fars	OM	OM	OM	OM	OM	OM	OM	OM
skin	OM	OM	OM	OM	OM	OM	OM	OM

datasets	5-fold CV		ELOO		ELOO-BR	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
thyroid	98.72%	4761.78	97.86%	13150.79	97.76%	13105.28
mushrooms	99.98%	7548.01	99.72%	19064.64	99.70%	19146.48
coil2000	93.88%	12747.31	94.03%	30909.36	93.92%	30584.37
penbased	99.71%	16644.09	93.73%	39055.87	95.35%	39237.54
rcv	NA	NA	NA	NA	NA	NA
adult	NA	NA	NA	NA	NA	NA
w8a	NA	NA	NA	NA	NA	NA
cod-rna	NA	NA	NA	NA	NA	NA
fars	OM	OM	OM	OM	OM	OM
skin	OM	OM	OM	OM	OM	OM

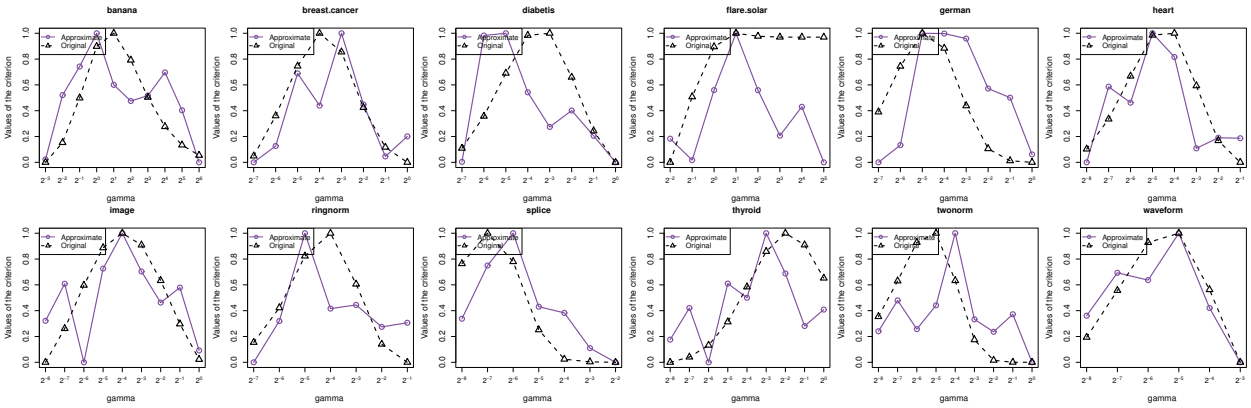

Figure 1: Comparison of the values of $\mathrm{KTA}(\mathbf{K})$ and $\mathrm{KTA}(\boldsymbol{v})$ for different kernel parameters.

more efficient than the baseline criteria while preserving similar classification accuracy. For the datasets containing about 10,000 examples, our criteria are around 400 times faster than 5-fold CV, and for the larger datasets, all criteria of $O(l^3)$ complexity will run for more than 120 hours. For the datasets with more than 100,000 examples, the memory cost required to store the kernel matrix has stopped all existing criteria from successfully running. According to the z statistic, there is no difference on accuracies between our criteria and the best of the compared criteria at the 95% level of significance.

Table 2: Comparison of the accuracy and time (seconds) among the randomized criterion on MCM $\mathrm{KTA}(\boldsymbol{v})$, the criterion on the direct stack of random features $\mathrm{KTA}(\mathbf{Z})$, and 5-fold CV with random Fourier features 5-CV-RFF.

datasets ($D = 100$)	$\mathrm{KTA}(\boldsymbol{v})$		$\mathrm{KTA}(\mathbf{Z})$		5-CV-RFF	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
thyroid	98.36%	1.75	98.72%	3.16	98.71%	99.27
mushrooms	98.72%	4.72	92.34%	3.60	90.40%	90.40
coil2000	93.88%	11.02	93.44%	4.28	93.61%	75.38
penbased	99.11%	3.74	99.24%	4.67	99.27%	77.29
datasets ($D = 500$)	$\mathrm{KTA}(\boldsymbol{v})$		$\mathrm{KTA}(\mathbf{Z})$		5-CV-RFF	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
thyroid	98.55%	11.27	98.76%	10.83	98.57%	1414.46
mushrooms	98.83%	15.32	93.47%	14.89	90.90%	1431.69
coil2000	93.75%	20.96	93.63%	17.51	94.01%	672.49
penbased	99.32%	12.73	99.38%	11.41	99.64%	921.53

In addition, the larger the dataset, the more efficiency gain the randomized criteria have, which is in agreement with the complexity analysis.

In the second experiment, we take KTA as an example to study the difference between the randomized criterion on $\mathbf{U}_{\boldsymbol{m}}$ and the criterion on the direct stack matrix $\mathbf{Z} = [z_0 \dots z_l] \in \mathbb{R}^{D\times l}$ with $z_i = [\cos(x_i^{\mathrm{T}} w_0 + b_0), \dots, \cos(x_i^{\mathrm{T}} w_{D-1} + b_{D-1})] \in \mathbb{R}^D$. $\mathcal{C}_2(\boldsymbol{v})$ is the randomized criterion for KTA on $\mathbf{U}_{\boldsymbol{m}}$. We rewrite $\mathcal{C}_2(\boldsymbol{v}) = \mathrm{KTA}(\boldsymbol{v})$. The criterion on $\mathbf{Z}$ can be defined as $\mathrm{KTA}(\mathbf{Z}) = \frac{(\mathbf{Z}\boldsymbol{y})^{\mathrm{T}}\mathbf{Z}\boldsymbol{y}}{l\|\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\|_{\mathrm{F}}}$. For completeness, we also compare 5-fold CV with random Fourier features (5-CV-RFF). In each fold of 5-CV-RFF, we solve linear ridge regression with random features (Rahimi and Recht 2008). The results are shown in Table 2. It is worth noting that this experiment is conducted on a PC, while the first experiment is conducted on our computing cluster. As expected, $\mathrm{KTA}(\boldsymbol{v})$ and $\mathrm{KTA}(\mathbf{Z})$ are much faster than 5-CV-RFF. In general, the efficiencies of $\mathrm{KTA}(\mathbf{Z})$ and $\mathrm{KTA}(\boldsymbol{v})$ are close to each other. However, for the mushrooms dataset, the accuracy of $\mathrm{KTA}(\mathbf{Z})$ for $D = 500$ is significantly lower than that of $\mathrm{KTA}(\boldsymbol{v})$, and even lower than $\mathrm{KTA}(\boldsymbol{v})$ for $D = 100$. There are two reasons for this. First, $\mathrm{KTA}(\boldsymbol{v})$ converges faster than $\mathrm{KTA}(\mathbf{Z})$, implying that a bigger D is required for $\mathrm{KTA}(\mathbf{Z})$. Second, $\mathrm{KTA}(\boldsymbol{v})$ adopts the spectrum of the labeled MCM and hence involves more data information for kernel selection. We could use $\mathbf{Z}$ to conduct randomized kernel selection in practice, but we need to sample much more random vectors to guarantee satisfactory performance.

The third experiment takes $\mathrm{KTA}(\mathbf{K})$ and $\mathrm{KTA}(\boldsymbol{v})$ as examples to gain deep insights on the difference between the original and randomized criteria. To show the general applicability of our method, this experiment adopts another set of smaller but widely-used benchmark datasets[6] in the model selection community (Rätsch, Onoda, and Müller 2001; Chapelle et al. 2002; Cawley and Talbot 2010). We check the values of $\mathrm{KTA}(\mathbf{K})$ and $\mathrm{KTA}(\boldsymbol{v})$ for different kernel parameters. The results are shown in Figure 1, in which we can find that the values of kernel parameters that reach the highest points of the curves for $\mathrm{KTA}(\mathbf{K})$ and $\mathrm{KTA}(\boldsymbol{v})$ are the same or very close (please pay attention to the x-axis), which means that the optimal kernels selected by maximizing $\mathrm{KTA}(\mathbf{K})$ and $\mathrm{KTA}(\boldsymbol{v})$ are the same or very close.

[6]http://theoval.cmp.uea.ac.uk/matlab/

Conclusions

In this paper, we specifically designed a kind of randomized MCMs for kernel selection and established the connection between the kernel functions and the spectra of the randomized MCMs. We provided the theoretical insights indicating that kernel selection on the spectra of the randomized MCMs is rational. Under the guarantee of the theoretical results, we defined a series of randomized kernel selection criteria with the spectra of the MCMs, which are of log-linear time complexity and linear space complexity. This kind of scalability alleviates the computational bottleneck faced by existing kernel selection approaches. We empirically verified the effectiveness of the randomized criteria and provided some deep insights of the randomized criteria.

Acknowledgments

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3007-01-01 and BAS/1/1624-01-01, National Natural Science Foundation of China (No. 61170019) and National Natural Science Foundation of China (No. 61673293).

References

Anguita, D.; Ghio, A.; Oneto, L.; and Ridella, S. 2012. In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems* 23(9):1390–1406.

Bach, F. 2013. Sharp analysis of low-rank kernel matrix approximations. In *COLT 2013*, 185–209.

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Bartlett, P. L.; Boucheron, S.; and Lugosi, G. 2002. Model selection and error estimation. *Machine Learning* 48(1–3):85–113.

Bochner, S. 1933. Monotone funktionen, Stieltjessche integrale und harmonische analyse. *Mathematische Annalen* 108(1):378–410.

Cawley, G. C., and Talbot, N. L. 2007. Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research* 8:841–861.

Cawley, G. C., and Talbot, N. L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11:2079–2107.

Chapelle, O.; Vapnik, V.; Bousquet, O.; and Mukherjee, S. 2002. Choosing multiple parameters for support vector machines. *Machine Learning* 46(1–3):131–159.

Cheng, Y.; Yu, F. X.; Feris, R. S.; Kumar, S.; Choudhary, A.; and Chang, S.-F. 2015. An exploration of parameter redundancy in deep networks with circulant projections. In *ICCV 2015*, 2857–2865.

Cortes, C.; Kloft, M.; and Mohri, M. 2013. Learning kernels using local Rademacher complexity. In *NIPS 26*. 2760–2768.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2009. L_2 regularization for learning kernels. In *UAI 2009*, 109–116.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2010. Two-stage learning kernel algorithms. In *ICML 2010*, 239–246.

Cristianini, N.; Shawe-Taylor, J.; Elisseeff, A.; and Kandola, J. S. 2002. On kernel-target alignment. In *NIPS 14*. 367–373.

Cucker, F., and Smale, S. 2002. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society* 39(1):1–49.

Ding, L., and Liao, S. 2014a. Approximate consistency: Towards foundations of approximate kernel selection. In *ECML 2014*, 354–369. Springer, Berlin.

Ding, L., and Liao, S. 2014b. Model selection with the covering number of the ball of RKHS. In *CIKM 2014*, 1159–1168.

Ding, L., and Liao, S. 2017. An approximate approach to automatic kernel selection. *IEEE Transactions on Cybernetics* 47(3):554–565.

Golub, G. H.; Heath, M.; and Wahba, G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–223.

Gray, R. 1972. On the asymptotic eigenvalue distribution of Toeplitz matrices. *IEEE Transactions on Information Theory* 18(6):725–730.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012a. A kernel two-sample test. *Journal of Machine Learning Research* 13(1):723–773.

Gretton, A.; Sejdinovic, D.; Strathmann, H.; Balakrishnan, S.; Pontil, M.; Fukumizu, K.; and Sriperumbudur, B. K. 2012b. Optimal kernel choice for large-scale two-sample tests. In *NIPS 25*, 1205–1213.

Hamid, R., and Xiao, Y. 2014. Compact random feature maps. In *ICML 2014*, 19–27.

Hinrichs, A., and Vybíral, J. 2011. Johnson-lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms* 39(3):391–398.

Kar, P., and Karnick, H. 2012. Random feature maps for dot product kernels. In *AISTATS 2012*, 583–591.

Kimeldorf, G. S., and Wahba, G. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* 41(2):495–502.

Le, Q.; Sarlòs, T.; and Smola, A. 2013. Fastfood — approximating kernel expansions in loglinear time. In *ICML 2013*, 244–252.

Liu, Y., and Liao, S. 2015. Eigenvalues ratio for kernel selection of kernel methods. In *AAAI 2015*, 2814–2820.

Liu, Y.; Liao, S.; Lin, H.; Yue, Y.; and Wang, W. 2017. Infinite kernel learning: generalization bounds and algorithms. In *AAAI 2017*.

Liu, Y.; Jiang, S.; and Liao, S. 2014. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In *ICML 2014*, 324–332.

Luntz, A., and Brailovsky, V. 1969. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica* 3(6):6–12.

Mahoney, M. W., and Drineas, P. 2009. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences* 106(3):697–702.

Nguyen, C. H., and Ho, T. B. 2007. Kernel matrix evaluation. In *IJCAI 2007*, 987–992.

Rahimi, A., and Recht, B. 2008. Random features for large-scale kernel machines. In *NIPS 20*. 1177–1184.

Rahimi, A., and Recht, B. 2009. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NIPS 22*. 1313–1320.

Rätsch, G.; Onoda, T.; and Müller, K. R. 2001. Soft margins for AdaBoost. *Machine Learning* 42(3):287–320.

Schoenberg, I. J. 1942. Positive definite functions on spheres. *Duke Mathematical Journal* 9(1):96–108.

Singleton, R. C. 1969. An algorithm for computing the mixed radix fast Fourier transform. *IEEE Transactions on Audio and Electroacoustics* 17(2):93–103.

Song, G., and Xu, Y. 2010. Approximation of high-dimensional kernel matrices by multilevel circulant matrices. *Journal of Complexity* 26(4):375–405.

Song, L.; Smola, A.; Gretton, A.; Bedo, J.; and Borgwardt, K. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research* 13:1393–1434.

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Lanckriet, G. R.; and Schölkopf, B. 2009. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS 22*, 1750–1758.

Tyrtyshnikov, E. E. 1996. A unifying approach to some old and new theorems on distribution and clustering. *Linear Algebra and its Applications* 232:1–43.

Vedaldi, A., and Zisserman, A. 2012. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(3):480–492.

Vybíral, J. 2011. A variant of the Johnson-Lindenstrauss lemma for circulant matrices. *Journal of Functional Analysis* 260(4):1096–1105.

Yang, J.; Sindhwani, V.; Avron, H.; and Mahoney, M. 2014. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *ICML 2014*, 485–493.

Yu, F.; Kumar, S.; Gong, Y.; and Chang, S.-F. 2014. Circulant binary embedding. In *ICML 2014*, 946–954.

Zhang, T. 2005. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation* 17(9):2077–2098.