Hypothesis Sketching for Online Kernel Selection in Continuous Kernel Space

Xiao Zhang and **Shizhong Liao***
College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
szliao@tju.edu.cn

Abstract

Online kernel selection in continuous kernel space is more complex than that in discrete kernel set. But existing online kernel selection approaches for continuous kernel spaces have linear computational complexities at each round with respect to the current number of rounds and lack sublinear regret guarantees due to the continuously many candidate kernels. To address these issues, we propose a novel hypothesis sketching approach to online kernel selection in continuous kernel space, which has constant computational complexities at each round and enjoys a sublinear regret bound. The main idea of the proposed hypothesis sketching approach is to maintain the orthogonality of the basis functions and the prediction accuracy of the hypothesis sketches in a time-varying reproducing kernel Hilbert space. We first present an efficient dependency condition to maintain the basis functions of the hypothesis sketches under a computational budget. Then we update the weights and the optimal kernels by minimizing the instantaneous loss of the hypothesis sketches using the online gradient descent with a compensation strategy. We prove that the proposed hypothesis sketching approach enjoys a regret bound of order $O(\sqrt{T})$ for online kernel selection in continuous kernel space, which is optimal for convex loss functions, where T is the number of rounds, and reduces the computational complexities at each round from linear to constant with respect to the number of rounds. Experimental results demonstrate that the proposed hypothesis sketching approach significantly improves the efficiency of online kernel selection in continuous kernel space while retaining comparable predictive accuracies.

1 Introduction

Online kernel learning obtains hypotheses incrementally in a reproducing kernel Hilbert space in a single-pass over the data, which aims to make a sequence of accurate predictions [Freund and Schapire, 1999; Kivinen *et al.*, 2001; Crammer *et al.*, 2003; Zhang and Liao, 2019]. Since the performance of online kernel learning significantly depends on the chosen kernel, kernel selection is critical to online kernel learning. In the literature of online kernel learning, the kernel function is typically chosen beforehand, which is not theoretically sound and needs multiple passes over all the data. In batch learning settings, most of the offline kernel selection approaches, such as cross-validation [Liu *et al.*, 2014; Liu *et al.*, 2018] and randomized kernel selection [Ding *et al.*, 2018], first learn the hypotheses on the training data for each candidate kernel, and then select the optimal kernel on the validation data. But in online learning settings, there is no such delineation among training, validation and testing [Diethe and Girolami, 2013; Zhang *et al.*, 2019]. The selection of optimal kernels for online kernel learning is more challenging than offline kernel selection, which only has access to the arrived data for kernel selection at each round, must construct the hypothesis space incrementally in low computational complexities, and requires a sublinear regret without i.i.d. assumption. Given candidate kernels, we refer to the dynamic selection of the optimal kernel at each round for online kernel learning as *online kernel selection*, which intermixes kernel selection and training, and requires a sublinear regret and constant computational complexities at each round. We call an online kernel selection problem using a continuous kernel space containing continuously many candidate kernels the *continuous online kernel selection* problem, and that using a discrete kernel set containing a finite number of candidate kernels the *discrete online kernel selection* problem.

Recently, several online kernel selection approaches have been presented. Foster et al. [2017] presented a meta-algorithm framework using the multi-scale aggregation for online model selection, which can be used for discrete online kernel selection. This meta-algorithm framework enjoys a sublinear regret bound that is suitable for multi-scale losses. Yang et al. [2012] formulated a randomized online kernel selection approach, which measures the relative importance of the candidate kernels by maintaining a probability distribution via an exponential weighted average [Auer *et al.*, 2002; Cesa-Bianchi and Lugosi, 2006]. This randomized online kernel selection approach to discrete online kernel selection has a quadratic overall time complexity and a linear space complexity with respect to the number of rounds. To address this issue, Zhang et al. [2018] proposed a random-

*Corresponding author

ized approach using a novel kernel alignment for discrete online kernel selection with a strongly convex loss function, which has constant computational complexities at each round. These two randomized approaches have sublinear regret bound for discrete online kernel selection, but cannot be applied to continuous online kernel selection with regret guarantees. Besides, they are not efficient for online kernel selection with a large number of candidate kernels. Adaptive kernel is another kind of online kernel selection approaches that use the gradient descent to minimize the losses for the updating of the hypotheses and the kernel parameters [Singh and Príncipe, 2011; Chen *et al.*, 2016; Nguyen *et al.*, 2017]. Although existing adaptive kernel approaches can be applied to continuous online kernel selection, they do not have sublinear regret guarantees that are essential for online kernel learning.

The aim of this paper is to propose a novel hypothesis sketching approach to continuous online kernel selection, which is theoretically solid and computationally efficient. We construct the hypothesis sketches by the basis function vector sketching and the weight vector sketching. Then we define the instantaneous loss of the hypothesis sketches, and use the gradient of this instantaneous loss to update the optimal kernel at each round. For regret analysis, we decompose the regret for continuous online kernel selection into three terms and derive a sublinear regret bound. When applied to continuous online kernel selection, our hypothesis sketching approach has only linear time and space complexities with respect to the budget at each round. Finally, we empirically demonstrate that our hypothesis sketching approach to online kernel selection achieves better performance than the existing approaches in efficiency with a comparable accuracy.

2 Notations and Background

Let $\mathbb{R}$ and $\mathbb{R}_+$ be the set of real numbers and the set of positive real numbers, respectively. Let $[T] = \{1, 2, \ldots, T\}$, $[\cdot]_+$ be the function satisfying $[a]_+ = a$ when $a > 0$ and $[a]_+ = 0$ when $a \leq 0$, $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^T \subseteq (\mathcal{X} \times \mathcal{Y})^T$ be the sequence of T instances, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. We denote the Moore-Penrose pseudoinverse of $\boldsymbol{A}$ by $\boldsymbol{A}^\dagger$, the loss function by $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{0\}$, the kernel function by $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and its corresponding kernel matrix by $\boldsymbol{K} = (\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)) \in \mathbb{R}^{T \times T}$. The reproducing kernel Hilbert space (RKHS) associated with κ is defined as $\mathcal{H}_\kappa = \overline{\text{span}\{\kappa(\cdot, \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}}$. We denote a set containing N candidate kernels for online kernel selection by $\mathcal{K}_N = \{\kappa_i\}_{i=1}^N$.

Given a candidate kernel set $\mathcal{K}_N = \{\kappa_i\}_{i=1}^N$, online kernel selection is to update the hypothesis and select the optimal kernel from $\mathcal{K}_N$ at each round in an online setting, which requires a sublinear regret guarantee and low computational complexities. Given convex loss functions, Yang et al. [2012] proposed a randomized kernel selection approach to discrete online kernel selection, which enjoys the regret bound of order $O(N\sqrt{T})$ against the optimal hypothesis in hindsight

$$g^* = \mathop{\arg\min}_{f \in \mathcal{H}_{\kappa_i},\ \kappa_i \in \mathcal{K}_N} \sum_{t=1}^T \ell(f(\boldsymbol{x}_t), y_t).$$

But the regret bound of order $O(N\sqrt{T})$ for online kernel selection is invalid for continuous online kernel selection due to the unbounded regret when $N \to \infty$. In this paper, we propose an efficient hypothesis sketching approach to online kernel selection, which enjoys a sublinear regret bound for continuous online kernel selection.

3 Novel Hypothesis Sketching Approach

In this section, we propose a novel hypothesis sketching approach, which has an efficient computation and can be applied to continuous online kernel selection with regret guarantees. We first define the hypothesis sketches as follows. Let

$$\boldsymbol{\psi}_\kappa^{(t)}(\cdot) = [\kappa(\cdot, \tilde{\boldsymbol{x}}_1), \kappa(\cdot, \tilde{\boldsymbol{x}}_2), \ldots, \kappa(\cdot, \tilde{\boldsymbol{x}}_{|\mathcal{V}_t|})]^\intercal, \quad \tilde{\boldsymbol{x}}_i \in \mathcal{V}_t,$$

be the *basis function vector*, where $\mathcal{V}_t = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{|\mathcal{V}_t|}\}$ is the buffer of examples at round t and $|\mathcal{V}_t|$ is the size of $\mathcal{V}_t$, and

$$\boldsymbol{\omega}^{(t)} = \left[\omega_1^{(t)}, \omega_2^{(t)}, \ldots, \omega_{|\mathcal{V}_t|}^{(t)}\right]^\intercal$$

be the *weight vector*. Then we call

$$f_t(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_\kappa^{(t)}(\cdot) \rangle$$

the *hypothesis sketch* for prediction at round t, which can be seen as an approximation of the original hypothesis over the first $t-1$ instances[1]:

$$h_t(\cdot) = \sum_{i \in [t-1]} \beta_{t,i}^* \kappa(\cdot, \boldsymbol{x}_i),$$

where $\{\beta_{t,i}^*\}_{i \in [t-1]}$ is the set of the optimal coefficients of the original hypothesis. Then we give the novel hypothesis sketching approach under a fixed budget. Specifically, we restrict the size of the hypothesis sketch $f_t(\cdot)$ by $|\mathcal{V}_t| \leq B$ at each round, where the budget $B > 0$. The novel hypothesis sketching approach includes the following two steps: the basis function vector sketching and the weight vector sketching.

Basis Function Vector Sketching: We first define a criterion to maintain the basis function vector of the hypothesis sketch, called randomized linear dependency condition. At round t, we use the inner product to measure the correlation between the newly arrived kernel function $\kappa(\cdot, \boldsymbol{x}_t)$ and the j-th kernel function in the basis function vector as follows:

$$q_j^{(t)} = \langle \kappa(\cdot, \boldsymbol{x}_t), \kappa(\cdot, \tilde{\boldsymbol{x}}_j) \rangle_{\mathcal{H}_\kappa}, \quad \tilde{\boldsymbol{x}}_j \in \mathcal{V}_t,$$

which yields a probability vector

$$\boldsymbol{p}_t = \left[q_1^{(t)}, q_2^{(t)}, \ldots, q_{|\mathcal{V}_t|}^{(t)}\right]^\intercal \Big/ \sum_{i=1}^{|\mathcal{V}_t|} q_i^{(t)} \in \mathbb{R}_+^{|\mathcal{V}_t|}.$$

We regard $\boldsymbol{p}_t$ as a multinomial distribution in the probability simplex on $[|\mathcal{V}_t|]$, and sample s ($s \ll |\mathcal{V}_t|$) kernel functions without replacement from the basis function vector according to $\boldsymbol{p}_t$, where the indices of the sampled kernel functions are $k_j^{(t)} \in [|\mathcal{V}_t|], j = 1, 2, \ldots s$. Then the *Randomized Linear Dependency* (RLD) condition at round t is given by

$$\delta_t := \min_{\boldsymbol{\alpha}_t \in \mathbb{R}^s} \left\| \kappa(\cdot, \boldsymbol{x}_t) - \sum_{i=1}^s \alpha_i^{(t)} \kappa(\cdot, \tilde{\boldsymbol{x}}_{k_i^{(t)}}) \right\|_{\mathcal{H}_\kappa}^2 > \nu, \quad (1)$$

[1] It follows from the representer theorem that it is sufficient to consider the original hypothesis of the linear combination form.

where $\nu > 0$ is the RLD parameter and

$$\boldsymbol{\alpha}_t = \left[\alpha_1^{(t)}, \alpha_2^{(t)}, \ldots, \alpha_s^{(t)}\right]^\intercal \in \mathbb{R}^s.$$

RLD condition measures the linear dependency between the newly arrived kernel function and the random subset of the kernel functions in the basis function vector. Solving the minimization problem in (1) yields the optimum solution

$$\boldsymbol{\alpha}_t^* = \left(\widetilde{\boldsymbol{K}}_s^{(t)}\right)^\dagger \widetilde{\boldsymbol{\psi}}_s^{(t)} \in \mathbb{R}^s, \tag{2}$$

where

$$\widetilde{\boldsymbol{\psi}}_s^{(t)} = \left[\kappa(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_{k_1^{(t)}}), \ldots, \kappa(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_{k_s^{(t)}})\right]^\intercal \in \mathbb{R}^s,$$

and

$$\widetilde{\boldsymbol{K}}_s^{(t)} = [\kappa(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)]_{i,j=k_1^{(t)}}^{k_s^{(t)}} \in \mathbb{R}^{s\times s}.$$

Given a budget B, if $|\mathcal{V}_t| < B$ at round t, we directly insert the kernel function $\kappa(\cdot, \boldsymbol{x}_t)$ into the basis function vector. Next, we consider the case that the size of the basis function vector reaches the budget, i.e., $|\mathcal{V}_t| = B$. When RLD condition holds, we delete the kernel function $\kappa(\cdot, \tilde{\boldsymbol{x}}_{r_t})$ with the smallest weight

$$r_t = \operatorname*{arg\,min}_{i\in[|\mathcal{V}_t|]} \left|\omega_i^{(t)}\right|,$$

and insert $\kappa(\cdot, \boldsymbol{x}_t)$ into the basis function vector, which is equivalent to $\mathcal{V}_{t+1} = \mathcal{V}_t \setminus \{\tilde{\boldsymbol{x}}_{r_t}\} \cup \{\boldsymbol{x}_t\}$, otherwise set $\mathcal{V}_{t+1} = \mathcal{V}_t$. Finally, we can obtain the updated basis function vector

$$\boldsymbol{\psi}_\kappa^{(t+1)}(\cdot) = [\kappa(\cdot, \tilde{\boldsymbol{x}}_1), \ldots, \kappa(\cdot, \tilde{\boldsymbol{x}}_{|\mathcal{V}_{t+1}|})]^\intercal, \;\; \tilde{\boldsymbol{x}}_i \in \mathcal{V}_{t+1},$$

and the hypothesis sketch after the basis function vector sketching as $f_t^{\mathrm{b}}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_\kappa^{(t+1)}(\cdot)\rangle$.

Weight Vector Sketching: After the basis function vector sketching, we maintain the weight vector of the hypothesis sketch. Given a convex loss function $\ell(\cdot,\cdot)$, at round t, instead of the instantaneous loss of the original hypothesis, we define the *Sketched Instantaneous Loss* (SIL) of the hypothesis sketch $f_t^{\mathrm{b}}(\cdot)$ by

$$\mathrm{SIL}(f_t^{\mathrm{b}}(\boldsymbol{x}_t)) := \ell(f_t^{\mathrm{b}}(\boldsymbol{x}_t), y_t) = \ell(\langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_\kappa^{(t+1)}(\boldsymbol{x}_t)\rangle, y_t).$$

Then, given a stepsize $\eta > 0$, we update the weight of the hypothesis sketch $f_t^{\mathrm{b}}(\cdot)$ using the Kernelized Online Gradient Descent (KOGD) [Kivinen *et al.*, 2001] for SIL as follows:

$$f_t^{\mathrm{w}}(\cdot) = f_t^{\mathrm{b}}(\cdot) - \eta \nabla_{f_t^{\mathrm{b}}(\boldsymbol{x}_t)} \mathrm{SIL}(f_t^{\mathrm{b}}(\boldsymbol{x}_t))\, \kappa(\cdot, \boldsymbol{x}_t).$$

The key ingredient of the weight vector sketching is to compensate the weight vector when ignoring the newly arrived examples. When $|\mathcal{V}_t| < B$, we obtain $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{\boldsymbol{x}_t\}$ and insert the following weight into the weight vector

$$\omega_{|\mathcal{V}_{t+1}|}^{(t+1)} = -\eta \nabla_{f_t^{\mathrm{b}}(\boldsymbol{x}_t)} \mathrm{SIL}(f_t^{\mathrm{b}}(\boldsymbol{x}_t)).$$

When the budget has been reached, we discuss the weight vector sketching in two cases: (a) for the case that RLD condition does not hold, we compensate the weight vector as

$$\omega_i^{(t+1)} = \omega_i^{(t)} - \eta \alpha_t^*(i) \nabla_{f_t^{\mathrm{b}}(\boldsymbol{x}_t)} \mathrm{SIL}(f_t^{\mathrm{b}}(\boldsymbol{x}_t)),$$

where $i \in \{k_1^{(t)}, \ldots, k_s^{(t)}\}$ and $\alpha_t^*(i)$ is the i-th component of $\boldsymbol{\alpha}_t^*$ in (2); (b) for the case that RLD condition holds, we replace $\kappa(\cdot, \tilde{\boldsymbol{x}}_{r_t})$ with $\kappa(\cdot, \boldsymbol{x}_t)$ in the basis function vector, and obtain the weight corresponding to $\kappa(\cdot, \boldsymbol{x}_t)$ as

$$\omega_{r_t}^{(t+1)} = -\eta \nabla_{f_t^{\mathrm{b}}(\boldsymbol{x}_t)} \mathrm{SIL}(f_t^{\mathrm{b}}(\boldsymbol{x}_t)).$$

We finally obtain the hypothesis sketch after weight vector sketching as

$$f_t^{\mathrm{w}}(\cdot) = \langle \boldsymbol{\omega}^{(t+1)}, \boldsymbol{\psi}_\kappa^{(t+1)}(\cdot)\rangle.$$

4 Application to Online Kernel Selection

In this section, we apply the proposed hypothesis sketching approach to continuous online kernel selection. Specifically, we update the optimal kernel at each round to determine the reproducing kernel Hilbert space (RKHS) in which the prediction and the hypothesis sketching are performed at the next round. In this paper, we focus on the continuous kernel space containing differentiable candidate kernels, denoted by $\mathcal{K}_\Omega = \{\kappa_\sigma |\ \sigma \in \Omega\}$, where Ω is the kernel parameter space of $\mathcal{K}_\Omega$ and κ_σ is the kernel function with the kernel parameter $\sigma \in \Omega$. For convenience, in the following sections, we denote $\boldsymbol{\psi}_{\kappa_\sigma}^{(t+1)}(\cdot)$ by $\boldsymbol{\psi}_\sigma^{(t+1)}(\cdot)$. At round t, we select the optimal kernel σ_{t+1} by minimizing SIL of the hypothesis sketch $f_t^{\mathrm{w}}(\cdot) = \langle \boldsymbol{\omega}^{(t+1)}, \boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot)\rangle$ as follows:

$$\sigma_{t+1} = \operatorname*{arg\,min}_{\sigma_t\in\Omega} \mathrm{SIL}(f_t^{\mathrm{w}}(\boldsymbol{x}_t)). \tag{3}$$

When the buffer $\mathcal{V}_t$ is changed, i.e., $\mathcal{V}_{t+1} \neq \mathcal{V}_t$, we update the optimal kernel by minimizing (3) using the online gradient descent [Zinkevich, 2003] as follows:

$$\sigma_{t+1} = \sigma_t - \rho\, \nabla_{\sigma_t} \mathrm{SIL}(f_t^{\mathrm{w}}(\boldsymbol{x}_t)), \tag{4}$$

where $\rho > 0$ is the stepsize. After the hypothesis sketching and the optimal kernel updating at round t, we obtain the hypothesis sketch for prediction at round $t+1$ as follows:

$$f_{t+1}(\cdot) = \langle \boldsymbol{\omega}^{(t+1)}, \boldsymbol{\psi}_{\sigma_{t+1}}^{(t+1)}(\cdot)\rangle.$$

Next, we specify the loss function and the candidate kernels. For a hinge loss function $\ell\left(f(\boldsymbol{x}), y\right) = [1 - yf(\boldsymbol{x})]_+$, $\nabla_{f(\boldsymbol{x})}\mathrm{SIL}(f(\boldsymbol{x}))$ is $-y$ if $yf(\boldsymbol{x}) < 1$ and 0 otherwise. For a candidate Gaussian kernel function[2]

$$\kappa_\sigma(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2/(2\sigma^2)\right), \sigma > 0,$$

we set $\gamma = 1/(2\sigma^2)$ and denote this Gaussian kernel function by $\kappa_\gamma(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\gamma\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2\right)$. Then, (4) is equivalent to

$$\gamma_{t+1} = \gamma_t - \rho y_t \sum_{j=1}^{|\mathcal{V}_{t+1}|} \omega_j^{(t+1)} \kappa_{\gamma_t}(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_j)\|\boldsymbol{x}_t - \tilde{\boldsymbol{x}}_j\|^2.$$

We finally summarize the above stages into Algorithm 1, called OKS-SIL.

5 Theoretical Analysis

In this section, we prove the regret bound of our hypothesis sketching approach to continuous online kernel selection, analyze the computational complexities of our approach, and compare our approach with the existing online kernel selection approaches with respect to the theoretical results.

[2]The optimal kernel updating in (4) can be directly applied to different types of candidate kernels, including the anisotropic RBF kernels and other differentiable kernels.

Algorithm 1 OKS-SIL Algorithm

Require: The stepsizes η and ρ, budget B, RLD parameter ν, sampling size s, initial kernel parameter σ_1

Initialize the weight vector $\boldsymbol{\omega}^{(1)} = \mathbf{0}$ and the buffer $\mathcal{V}_1 = \emptyset$
for $t = 1, \ldots, T$ **do**
  Compute $\boldsymbol{\psi}_{\sigma_t}^{(t)}(\boldsymbol{x}_t) = [\kappa_{\sigma_t}(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_1), \ldots, \kappa_{\sigma_t}(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_{|\mathcal{V}_t|})]^\intercal$
  Compute $f_t(\boldsymbol{x}_t) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t)}(\boldsymbol{x}_t) \rangle$
  Predict $\hat{y}_t = \operatorname{sgn}(f_t(\boldsymbol{x}_t))$
  if $y_t f_t(\boldsymbol{x}_t) < 1$ **then**
    if $|\mathcal{V}_t| < B$ **then**
      $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{\boldsymbol{x}_t\}$
    else
      if RLD condition holds **then**
        Set $r_t = \arg\min_{i \in [|\mathcal{V}_t|]} |\omega_i^{(t)}|$
        $\mathcal{V}_{t+1} = \mathcal{V}_t \setminus \{\tilde{\boldsymbol{x}}_{r_t}\} \cup \{\boldsymbol{x}_t\}$
      else
        $\mathcal{V}_{t+1} = \mathcal{V}_t$
      end if
    end if
    Update $\boldsymbol{\omega}^{(t+1)}$ using KOGD with the compensation
    if $\mathcal{V}_{t+1} \neq \mathcal{V}_t$ **then**
      Update the optimal kernel using $\gamma_{t+1} = \gamma_t - \rho y_t \sum_{j=1}^{|\mathcal{V}_{t+1}|} \omega_j^{(t+1)} \kappa_{\gamma_t}(\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_j) \|\boldsymbol{x}_t - \tilde{\boldsymbol{x}}_j\|^2$
    end if
  end if
end for

5.1 Regret Analysis

Consider a continuous kernel space $\mathcal{K}_\Omega$ containing candidate Gaussian kernels with the kernel parameter space $\Omega = \{\sigma | \sigma \in [\sigma_{\min}, \sigma_{\max}]\}$, where $\sigma_{\min}, \sigma_{\max} \in \mathbb{R}_+$. For convenience, we denote the Gaussian RKHS associated with the kernel parameter σ by $\mathcal{H}_\sigma$ in the following theoretical analysis. In contrast to the regret analysis for online kernel learning in a given hypothesis space, the key challenge of the regret analysis for continuous online kernel selection is to bound the regret of a hypothesis sequence in an incrementally constructed hypothesis space. Our goal is to analyze the regret against the optimal hypothesis f^* with the optimal kernel κ_{σ^*} as follows:

$$(f^*, \sigma^*) = \operatorname*{arg\,min}_{f \in \mathcal{H}_\sigma,\ \sigma \in \Omega} \sum_{t=1}^{T} \ell(f(\boldsymbol{x}_t), y_t),$$

where $f^*(\cdot) = \langle \boldsymbol{\omega}^*, \boldsymbol{\psi}_{\sigma^*}^*(\cdot) \rangle$. Since Gaussian RKHS has a nested structure [Minh, 2010; Yukawa, 2015] as $\mathcal{H}_{\sigma_{\max}} \subseteq \mathcal{H}_\sigma \subseteq \mathcal{H}_{\sigma_{\min}}, \sigma \in \Omega$, it is reasonable to assume that

$$\|f\|_{\mathcal{H}_{\sigma_{\min}}} \leq R, \quad f \in \mathcal{H}_\sigma,\ \sigma \in \Omega.$$

Let $\bar{f}_{\sigma_t}^*(\cdot) = \langle \boldsymbol{\omega}^*, \boldsymbol{\psi}_{\sigma_t}^*(\cdot) \rangle$ and denote the optimal hypothesis in hindsight in $\mathcal{H}_{\sigma_t}$ by $f_{\sigma_t}^*$. Finally, we demonstrate the regret bound of the proposed hypothesis sketching approach to continuous online kernel selection as in Theorem 1.

Theorem 1 (Regret Bound)**.** *Let $\ell(\cdot, \cdot)$ be a hinge loss function, $\mathcal{K}_\Omega$ be a continuous kernel space containing Gaussian kernels with the kernel parameter space $\Omega = \{\sigma | \sigma \in [\sigma_{\min}, \sigma_{\max}]\}$, and $C(\nu)$ be a decreasing function of ν which is of order $O(\ln T)$ of magnitude. For any sequence of T instances $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^T \subseteq (\mathcal{X} \times \mathcal{Y})^T$, let $C_{\max} = \max_{i,j \in [T]} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, assume that $\max_{t \in [T]} |\nabla_{\sigma_t} f_t^{\mathrm{w}}(\boldsymbol{x}_t)| \leq L$ and $\sigma_{\max} < \min_{\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j \in \mathcal{V},\ \tilde{\boldsymbol{x}}_i \neq \tilde{\boldsymbol{x}}_j} \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j\| / \sqrt{3}, \mathcal{V} = \cup_{t=1}^T \mathcal{V}_t$. Then, let $\{f_t\}_{t=1}^T$ be the hypothesis sequence generated by Algorithm 1, we have*

$$\sum_{t=1}^{T} \mathrm{E}\left[\ell(f_t(\boldsymbol{x}_t), y_t) - \ell(f^*(\boldsymbol{x}_t), y_t)\right] \leq \overline{\mathrm{R}}_a + \overline{\mathrm{R}}_b + \overline{\mathrm{R}}_c,$$

where $W = \sigma_{\min}^{-3} C_{\max} \max_{t \in [T]} \left(\left|\bar{f}_{\sigma_t}^*(\boldsymbol{x}_t)\right| + |f_t^{\mathrm{w}}(\boldsymbol{x}_t)|\right)$,

$$\overline{\mathrm{R}}_a = \frac{R^2}{2\eta} + \eta O(T) + O\left(\sqrt{\nu} T + C(\nu)\right),$$

$$\overline{\mathrm{R}}_b = T \max_{t \in [T]} (|\bar{f}_{\sigma_t}^*(\boldsymbol{x}_t)| + |f_{\sigma_t}^*(\boldsymbol{x}_t)|),$$

$$\overline{\mathrm{R}}_c = \frac{(\sigma^*)^2}{2\rho} + \frac{(W+L)^2}{2} \rho T + 2\sigma_{\max} W T.$$

Proof Sketch. We first decompose the instantaneous regret at round t into the following three terms

$$\underbrace{\mathrm{Reg}_t(f_t, f_{\sigma_t}^*)}_{\text{Optimization error}} + \underbrace{\mathrm{Reg}_t(f_{\sigma_t}^*, \bar{f}_{\sigma_t}^*)}_{\text{Estimation error}} + \underbrace{\mathrm{Reg}_t(\bar{f}_{\sigma_t}^*, f^*)}_{\text{Approximation error}},$$

where $\mathrm{Reg}_t(a, b) := \ell(a(\boldsymbol{x}_t), y_t) - \ell(b(\boldsymbol{x}_t), y_t)$. Then we prove the upper bounds of the three terms by bounding the gradient errors of the hypothesis sketches, and obtain the final regret bound. □

Remark 1. *For online classification, it is reasonable to assume that the values of the hypotheses are $O(1/\sqrt{T})$ as in [Zhao* et al.*, 2012; Hu* et al.*, 2015], because it does not have any influence on the prediction when multiplying the weight vector by a factor of order $O(1/\sqrt{T})$. Besides, it is a reasonable assumption that $\mathcal{S}$ is compact in online learning [Cesa-Bianchi and Lugosi, 2006], which indicates that $C_{\max}$ is a positive constant. Then, for $\forall t \in [T]$, we assume that $\mathcal{S}$ is compact and $|f_{\sigma_t}^*(\boldsymbol{x}_t)|$, $|\bar{f}_{\sigma_t}^*(\boldsymbol{x}_t)|$, $|f_t^{\mathrm{w}}(\boldsymbol{x}_t)|$ are of the order of $O(1/\sqrt{T})$, which yields $\overline{\mathrm{R}}_b = O(\sqrt{T})$ and $W = O(1/\sqrt{T})$. Setting $\eta, \rho = O(1/\sqrt{T})$ and $\nu = O(1/T)$, we obtain a $O(\sqrt{T})$ regret bound for continuous online kernel selection that is optimal for convex loss functions and an online gradient descent approach [Shalev-Shwartz, 2011; Hazan, 2016].*

5.2 Complexity Analysis

At each round, for the basis function vector sketching, constructing the multinomial distribution has a linear time complexity with respect to the budget B, and computing the RLD condition is in $O(s^3)$ time complexity, where s is the sampling size and $s \ll B$. The time complexities of the weight vector sketching and the optimal kernel updating are $O(s)$ and $O(B)$, respectively. Therefore, the time complexity of OKS-SIL at each round is $O(B)$, where B is the budget of

Approach	Computational complexities			Regret guarantees	
	Time (round t)	Time (overall)	Space	Candidate kernels	Regret bound
OKS	$O(N+t)$	$O(T^2+NT)$	$O(T)$	N	$O\left(N\sqrt{T}\right)$
OKL-GD	$O(t)$	$O(T^2)$	$O(T)$	Continuously many	–
OKS-SIL	$O(B)$	$O(BT)$	$O(B)$	Continuously many	$O\left(\sqrt{T}\right)$

Table 1: Comparisons between the proposed OKS-SIL and OKS, OKL-GD for convex functions (T: the number of rounds, N: the number of candidate kernels; t: the index of the round; B: the budget, a constant; Candidate kernels: the number of candidate kernels; "–": not available).

the hypothesis sketches which is a constant. Since OKS-SIL only stores the hypothesis sketch under the budget B, the total space complexity of OKS-SIL is $O(B)$. Table 1 summarizes the theoretical results of our OKS-SIL, OKS [Yang *et al.*, 2012] and OKL-GD [Chen *et al.*, 2016]. From Table 1, we have the following observations: (a) our OKS-SIL reduces the time and space complexities at each round from linear to constant with respect to the number of rounds; (b) for continuous online kernel selection, in contrast to the existing approaches without regret guarantees, our OKS-SIL enjoys a sublinear regret bound.

6 Experiments

In this section, we present experimental results to demonstrate the effectiveness and efficiency of our OKS-SIL. We compared OKS-SIL algorithm with the following state-of-the-art online kernel selection algorithms: (a) **OKS** [Yang *et al.*, 2012] maintains a probability distribution for discrete online kernel selection, and updates the optimal kernel and the hypothesis at each round according to this distribution; (b) **OKL-GD** [Chen *et al.*, 2016] uses KOGD to minimize the squared loss for the hypothesis updating and the kernel parameter updating[3]; (c) Similar to the two-stage approach [Cortes *et al.*, 2012], the two-pass algorithm **Proj-KA** uses the first $m = \min\{T, 10^4\}$ instances to form a validation set, selects the kernel using the kernel alignment criterion [Kandola *et al.*, 2002] on the validation set, and runs Projectron [Orabona *et al.*, 2008] with the selected kernel. For each benchmark dataset[4] we merged the training and testing data into a single dataset. Experiments were conducted over 20 different random permutations of the datasets and all the algorithms were implemented in R 3.3.2 on a PC with 3.60 GHz Intel Core i7 CPU and 16GB memory. The mistake rate is used to evaluate the performance for classification, which is defined by $\sum_{t=1}^{T} I(y_t f_t(\boldsymbol{x}_t) < 0)/T \times 100$.

We chose a set of Gaussian kernels with kernel parameters $\sigma \in \{2^{-(i+1)/2}, i = [-12 : +1 : 12]\}$ as the discrete candidate kernel set for OKS and Proj-KA, and restricted the kernel parameter space of OKS-SIL and OKL-GD to the closed interval $\Omega = [2^{-6.5}, 2^{5.5}]$ for fair comparison. The index i of the initial kernel parameter was chosen in $[-12 : +1 : -6]$ using uniform sampling, for avoiding the vanishing of the initial gradient. For our OKS-SIL, we set $B = 150$ for the small datasets ($T < 10^4$), $B = 200$ for the other datasets ($T \geq 10^4$), and $\nu = 0.9$, $s = 3$ according to the empirical analysis in the subsection "Parameter Influence". We used a time-varying stepsize $\rho_t = 1/t$ at round t for optimal kernel updating, and tuned the stepsize of KOGD in a range $10^{[-5:+1:0]}$ for all the algorithms. For OKS, we ran the algorithm in a single pass over the data with the smoothing parameter $\delta = 0.2$ and the stepsize of weight updating $\eta = \sqrt{2(1-\delta)\ln N/NT}$ as in [Yang *et al.*, 2012]. For Projectron, we set the projection parameter $U = 1/4\sqrt{(B+1)/\log(B+1)}$ as in [Orabona *et al.*, 2008].

6.1 Performance Evaluation

In this subsection, we evaluate the performances of OKS-SIL compared with OKL-GD, OKS, and Proj-KA for online classification on benchmark datasets. The results are listed in Table 2, including mistake rates (mean $\pm$ std) and running time after T rounds. From the results, we draw some observations as follows. First, the experimental results show that OKS-SIL is computationally more efficient than the other online kernel selection algorithms for online classification task when the number of rounds exceeds 1000. This is because the computation of the instantaneous loss on hypothesis sketches has a linear time complexity with respect to the budget per round, in comparison with the quadratic time complexity after T rounds of OKL-GD and OKS. Besides, the projection strategy of Projectron is in quadratic time complexity with respect to the budget B, while RLD condition in our hypothesis sketching approach is more efficient with a $O(s^3)$ time complexity, where $s \ll B$. Second, OKS-SIL has lower mistake rates than other algorithms on most datasets for online classification, and preserves a comparable accuracy to the compared algorithms in terms of mistake rates on `w7a` and `cod-rna`, which demonstrates the effectiveness of OKS-SIL. The empirical performances of OKS-SIL conform to the regret analysis in Theorem 1. The main difference between OKL-GD and OKS-SIL is that, OKL-GD only updates the kernel parameter for the newly arrived kernel function, while OKS-SIL updates the kernel parameters for both the new kernel function and the previously chosen kernel functions, which yields better performances.

[3]For the online classification task, we use the hinge loss instead of the squared loss, and limit the number of instances in memory to 2000 for OKL-GD to prevent the curse of kernelization.

[4]http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Algorithm	german		svmguide3		spambase		mushrooms	
	Mistake (%)	Time (s)	Mistake (%)	Time (s)	Mistake (%)	Time (s)	Mistake (%)	Time (s)
OKS-SIL	**29.610 ± 0.363**	0.250	**21.480 ± 0.624**	**0.170**	**28.209 ± 0.104**	**1.526**	**0.351 ± 0.420**	**1.785**
OKL-GD	30.990 ± 0.899	**0.217**	23.821 ± 0.407	0.208	31.869 ± 0.145	5.567	3.957 ± 1.623	39.625
OKS	42.590 ± 1.049	0.253	29.879 ± 0.852	0.287	34.432 ± 0.282	3.965	7.749 ± 0.532	7.690
Proj-KA	38.280 ± 1.405	2.318	25.567 ± 0.829	3.307	30.596 ± 0.293	35.336	0.375 ± 0.415	175.815
Algorithm	a9a		w7a		ijcnn1		cod-rna	
	Mistake (%)	Time (s)	Mistake (%)	Time (s)	Mistake (%)	Time (s)	Mistake (%)	Time (s)
OKS-SIL	**23.753 ± 0.502**	**32.026**	2.975 ± 0.012	**72.826**	**9.574 ± 0.327**	**23.392**	12.989 ± 0.201	**89.190**
OKL-GD	23.929 ± 0.436	319.956	**2.973 ± 0.010**	434.908	9.582 ± 0.322	110.356	13.558 ± 0.225	109.410
OKS	24.302 ± 0.512	983.808	7.199 ± 0.860	837.220	9.873 ± 0.412	519.114	**12.927 ± 0.707**	1786.890
Proj-KA	23.901 ± 0.471	321.434	5.430 ± 1.816	622.044	11.940 ± 1.540	212.476	15.151 ± 0.304	163.144

Table 2: The mistake rates (Mistake) and running time (Time) of OKL-GD, OKS, Proj-KA and our OKS-SIL after T rounds for online classification task.

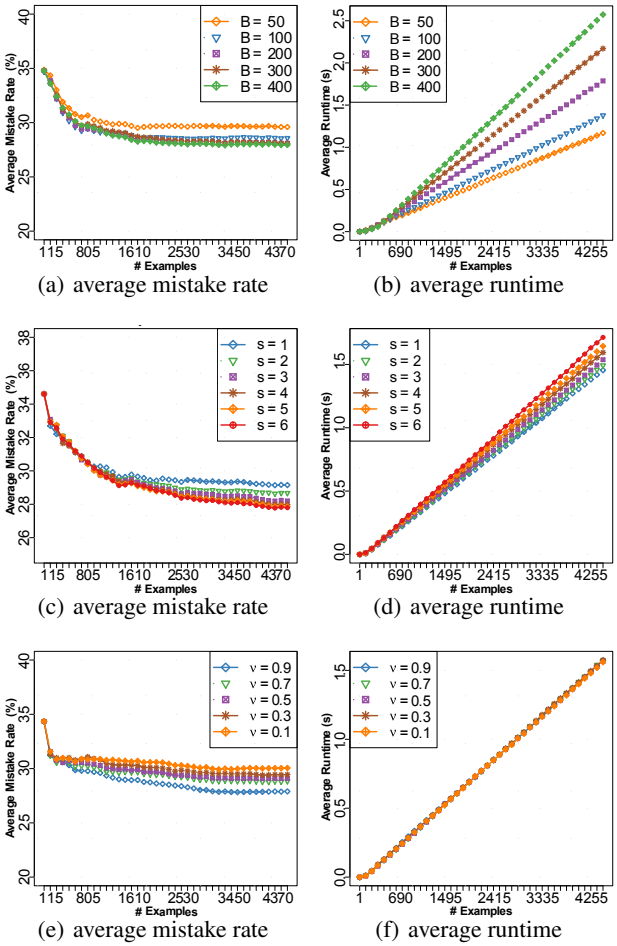
6.2 Parameter Influence

To analyze in more detail the behavior of OKS-SIL, we further conduct experiments with a range of values of B, s and ν on `spambase`. From the experimental results in Figure 1, we have the following observations for OKS-SIL.

For the budget $B \in \{50, 100, 200, 300, 400\}$ in Figure 1 (a)-(b), we observe that the mistake rate of OKS-SIL decreases gradually with the increase of the budget while the runtime increases. But when the budget $B > 50$, the varying B has little influence on the accuracy of OKS-SIL, which demonstrates the effectiveness of our hypothesis sketching approach in OKS-SIL. Figure 1 (c)-(d) show the performances of OKS-SIL using different sampling size $s \in \{1, 2, \ldots, 6\}$ in RLD condition. We can see that OKS-SIL with larger sampling sizes has better empirical performances in terms of the mistake rate, and the mistake rate of OKS-SIL is no longer significantly reduced when the sampling size is relatively large. For a larger s, more values of kernel functions need to be computed for RLD condition, leading to a higher running time cost. The RLD parameter ν controls a trade-off between $\sqrt{\nu}$ and $C(\nu)$ in $\overline{\mathrm{R}}_a$ in the regret bound of Theorem 1. From the results in terms of RLD parameter in Figure 1 (e)-(f), we can observe that OKS-SIL using RLD parameters closed to 1 yields better performances than that using other small RLD parameters, and OKS-SIL has a similar efficiency under different RLD parameters. The reason is that larger RLD parameters in (1) ensure that the kernel functions in the basis function vector are approximately orthogonal.

7 Conclusion

We have proposed a novel hypothesis sketching approach to online kernel selection in continuous kernel space, which is theoretically guaranteed and computationally efficient. Our sketching approach enjoys the optimal regret bound for online kernel selection, and has linear time and space complexities with respect to the budget at each round. The theoretical formulation and algorithmic implementation of the sketching mechanism are also promising for sketching approaches to online model selection and online learning.



(a) average mistake rate (b) average runtime

(c) average mistake rate (d) average runtime

(e) average mistake rate (f) average runtime

Figure 1: Average mistake rate and average runtime of OKS-SIL using different budget B, sampling size s and RLD parameter ν of RLD condition (1) on `spambase`.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 61673293).

References

[Auer *et al.*, 2002] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[Cesa-Bianchi and Lugosi, 2006] Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[Chen *et al.*, 2016] Badong Chen, Junli Liang, Nanning Zheng, and José C Príncipe. Kernel least mean square with adaptive kernel size. *Neurocomputing*, 191:95–106, 2016.

[Cortes *et al.*, 2012] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828, 2012.

[Crammer *et al.*, 2003] Koby Crammer, Jaz Kandola, and Yoram Singer. Online classification on a budget. In *Advances in Neural Information Processing Systems 16*, pages 225–232, 2003.

[Diethe and Girolami, 2013] Tom Diethe and Mark Girolami. Online learning with (multiple) kernels: A review. *Neural Computation*, 25(3):567–625, 2013.

[Ding *et al.*, 2018] Lizhong Ding, Shizhong Liao, Yong Liu, Peng Yang, and Xin Gao. Randomized kernel selection with spectra of multilevel circulant matrices. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 2910–2917, 2018.

[Foster *et al.*, 2017] Dylan J Foster, Satyen Kale, Mehryar Mohri, and Karthik Sridharan. Parameter-free online learning via model selection. In *Advances in Neural Information Processing Systems 30*, pages 6022–6032, 2017.

[Freund and Schapire, 1999] Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.

[Hazan, 2016] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3–4):157–325, 2016.

[Hu *et al.*, 2015] Junjie Hu, Haiqin Yang, Irwin King, Michael R Lyu, and Anthony Man-Cho So. Kernelized online imbalanced learning with fixed budgets. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2666–2672, 2015.

[Kandola *et al.*, 2002] J. Kandola, J. Shawe-Taylor, and N. Cristianini. On the extensions of kernel alignment. Technical Report Technical report 120, University of London, 2002.

[Kivinen *et al.*, 2001] Jyrki Kivinen, Alex J Smola, and Robert C Williamson. Online learning with kernels. In *Advances in Neural Information Processing Systems 14*, pages 785–792, 2001.

[Liu *et al.*, 2014] Yong Liu, Shali Jiang, and Shizhong Liao. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *Proceedings of the 31st International Conference on Machine Learning*, pages 324–332, 2014.

[Liu *et al.*, 2018] Yong Liu, Hailun Lin, Lizhong Ding, Weiping Wang, and Shizhong Liao. Fast cross-validation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2497–2503, 2018.

[Minh, 2010] Ha Quang Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

[Nguyen *et al.*, 2017] Tu Dinh Nguyen, Trung Le, Hung Bui, and Dinh Q. Phung. Large-scale online kernel learning with random feature reparameterization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2543–2549, 2017.

[Orabona *et al.*, 2008] Francesco Orabona, Joseph Keshet, and Barbara Caputo. The projectron: A bounded kernel-based perceptron. In *Proceedings of the 25th International Conference on Machine learning*, pages 720–727, 2008.

[Shalev-Shwartz, 2011] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011.

[Singh and Príncipe, 2011] Abhishek Singh and José C. Príncipe. Information theoretic learning with adaptive kernels. *IEEE Transactions on Signal Processing*, 91(2):203–213, 2011.

[Yang *et al.*, 2012] Tianbao Yang, Mehrdad Mahdavi, Rong Jin, Jinfeng Yi, and Steven C.H. Hoi. Online kernel selection: Algorithms and evaluations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 22–26, 2012.

[Yukawa, 2015] Masahiro Yukawa. Adaptive learning in Cartesian product of reproducing kernel Hilbert spaces. *IEEE Transactions on Signal Processing*, 63(22):6037–6048, 2015.

[Zhang and Liao, 2018] Xiao Zhang and Shizhong Liao. Online kernel selection via incremental sketched kernel alignment. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3118–3124, 2018.

[Zhang and Liao, 2019] Xiao Zhang and Shizhong Liao. Incremental randomized sketching for online kernel learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7394–7403, 2019.

[Zhang *et al.*, 2019] Xiao Zhang, Yun Liao, and Shizhong Liao. A survey on online kernel selection for online kernel learning. *WIREs Data Mining and Knowledge Discovery*, 9(2):e1295, 2019.

[Zhao *et al.*, 2012] Peilin Zhao, Jialei Wang, Pengcheng Wu, Rong Jin, and Steven C.H. Hoi. Fast bounded online gradient descent algorithms for scalable kernel-based online learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 169–176, 2012.

[Zinkevich, 2003] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.