Regret Bounds for Online Kernel Selection in Continuous Kernel Space

Xiao Zhang,[1,2] **Shizhong Liao,**[3] **Jun Xu,**[1,2*] **Ji-Rong Wen**[1,2]

[1] Gaoling School of Articial Intelligence, Renmin University of China
[2] Beijing Key Laboratory of Big Data Management and Analysis Methods
[3] College of Intelligence and Computing, Tianjin University
{zhangx89, junxu, jrwen}@ruc.edu.cn, szliao@tju.edu.cn

Abstract

Regret bounds of online kernel selection in a finite kernel set have been well studied, having at least an order $O(\sqrt{NT})$ of magnitude after T rounds, where N is the number of candidate kernels. But it is still an unsolved problem to achieve sublinear regret bounds of online kernel selection in a continuous kernel space under different learning frameworks. In this paper, to represent different learning frameworks of online kernel selection, we divide online kernel selection approaches in a continuous kernel space into two categories according to the order of selection and training at each round. Then we construct a surrogate hypothesis space that contains all the candidate kernels with bounded norms and inner products, representing the continuously varying hypothesis space. Finally, we decompose the regrets of the proposed online kernel selection categories into different types of instantaneous regrets in the surrogate hypothesis space, and derive optimal regret bounds of order $O(\sqrt{T})$ of magnitude under mild assumptions, independent of the cardinality of the continuous kernel space. Empirical studies verified the correctness of the theoretical regret analyses.

Introduction

In contrast to offline kernel selection (Ding et al. 2019; Liu et al. 2020), online kernel selection aims to select the optimal kernel for online kernel learning at each round, which conducts kernel selection and hypothesis training with regret guarantees in a single-pass over the data. Online kernel selection is one of the fundamental and critical problems of online kernel learning, for the kernels used determine the performance of online kernel learning. Existing offline kernel selection approaches cannot be directly applied to online kernel selection for the following two reasons: (1) there is no delineation among training, validation and testing phases in online learning (Diethe and Girolami 2013; Zhang, Liao, and Liao 2019; Muthukumar et al. 2019); (2) offline setting typically assumes that the data is generated independently and identically distributed, but the assumption is relaxed or eliminated in online settings (Rakhlin, Shamir, and Sridharan 2012).

*Corresponding author

Regret bounds of online kernel selection have been extensively studied for a *finite kernel set* containing a finite number of candidate kernels. Online kernel selection can be reduced to a problem of prediction with expert advice (Cesa-Bianchi and Lugosi 2006), where the finite kernel set corresponds to the set of experts, and predictions are obtained according to the weights and advice of experts at each round. Yang et al. (2012) presented an online approach to offline kernel selection with online-to-batch conversion, which has generalization guarantees. This online approach can be transformed into a randomized approach of online kernel selection with expert advice, which updates the weights using the exponential weighted average, trains hypotheses using the modified kernel perceptron. This randomized online kernel selection approach is in a linear time complexity at each round with respect to the current number of rounds and a linear space complexity with respect to the number of rounds. Foster et al. (2017) formulated an algorithm framework for online model selection with multi-scale expert advice, which enjoys tight regret bounds when the losses of the optimal hypotheses lie in different ranges. This framework can be applied to online kernel selection, where a min-max optimization problem needs to be solved. Another alternative strategy under the expert advice framework is online multiple kernel learning (Jin, Hoi, and Yang 2010; Nguyen 2017), which assigns weights to multiple hypothesis sequences corresponding to each candidate kernel, and predicts by combining the outputs of all the hypotheses at each round. But existing regret analyses for online kernel selection with expert advice are not suitable for online kernel selection in a continuous kernel space. More specifically, given a finite kernel set containing N candidate kernels, after T rounds, existing regret bounds of online kernel selection are at least of order $O(\sqrt{NT})$ against the best-in-hindsight hypothesis, which cannot be applied to a continuous kernel space due to the unbounded regrets when $N = \infty$.

Recently, Zhang and Liao (2018) presented a novel online kernel selection approach using incremental sketched kernel alignment, which enjoys an optimal regret bound independently of the number of candidate kernels. However, given N candidate kernels, this online kernel selection approach needs to maintain N kernel alignments at each round, having a linear time complexity with respect to N at each updating round, which is unfeasible when $N = \infty$. Adaptive kernel

approaches have been proposed for online kernel selection in a continuous kernel space, which simultaneously update the hypothesis and kernel parameter at each round using online gradient descent (Singh and Príncipe 2011; Chen et al. 2016). These adaptive kernel approaches require a linear space complexity and a quadratic overall time complexity with respect to the number of rounds. Nguyen et al. (2017) proposed an efficient adaptive kernel approach using random features (Rahimi and Recht 2007), which derives the gradient of the random Fourier features using a reparameterization trick, and optimizes the kernel parameter using online gradient descent. Although existing adaptive kernel approaches can be applied to online kernel selection in a continuous kernel space, they lack sublinear regret guarantees that are essential for online kernel selection. Zhang and Liao (2020) focused on the budgeted online kernel selection problem in a continuous kernel space, and proved a sublinear regret bound under the assumption that the budget maintenance function is of order $O(\ln T)$ of magnitude. But it is difficult to provide a lower bound of the budget, which may lead to high computational complexities of the online kernel selection process.

In this paper, we first define the hypothesis sketch sequence using the vectors of weight and basis with one buffer. Then we formulate two categories of online kernel selection with the hypothesis sketch sequence by considering the orders of selection and training at each round, in which online kernel selection has polylogarithmic computational complexities at each round with respect to the current number of rounds. We further derive the optimal regret bounds for the two online kernel selection categories in a continuous kernel space under mild assumptions. Finally, we empirically evaluate the performances of different categories of online kernel selection, verifying the correctness of the theoretical results.

Notations and Preliminaries

Let $[T] = \{1, 2, \ldots, T\}$, $\det(\boldsymbol{A})$ be the determinant of a nonsingular matrix $\boldsymbol{A}$, $\boldsymbol{A}^\dagger$ be the Moore-Penrose pseudoinverse of $\boldsymbol{A}$, $\mathcal{S} = \{\boldsymbol{z}_t\}_{t=1}^T$ be a sequence of T instances, where $\boldsymbol{z}_t = (\boldsymbol{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}^d$ is compact and $\mathcal{Y} = \mathbb{R}$ or $\{-1, 1\}$. We denote a convex loss function by $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ \cup \{0\}$, the gradient (or a subgradient) of $\ell(f(\boldsymbol{x}_t), y_t)$ at f by $\nabla_f \ell(f(\boldsymbol{x}_t), y_t)$, a kernel function by $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and the reproducing kernel Hilbert space (RKHS) associated with κ by $\mathcal{H}_\kappa = \overline{\operatorname{span}\{\kappa(\cdot, \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}}$. Let Ω be a parameter interval of candidate kernels, we define the *continuous kernel space* by $\mathcal{K}_\Omega = \{\kappa_\sigma \mid \sigma \in \Omega\}$, where σ is the kernel parameter of κ_σ.

Online kernel selection in a continuous kernel space $\mathcal{K}_\Omega$ aims to generate a hypothesis sequence $\{f_t\}_{t=1}^T \subseteq \bigcup_{\kappa_\sigma \in \mathcal{K}_\Omega} \mathcal{H}_{\kappa_\sigma}$ such that

$$\begin{aligned}\operatorname{Reg}_T(\{f_t\}_{t=1}^T, f^*) &:= \sum_{t=1}^T \left[\ell\left(f_t(\boldsymbol{x}_t), y_t\right) - \ell\left(f^*(\boldsymbol{x}_t), y_t\right)\right] \\ &= o(T),\end{aligned}$$

where f^* is a competing hypothesis that is typically defined as the best-in-hindsight hypothesis

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

We call $\operatorname{Reg}_T(\{f_t\}_{t=1}^T, f^*)$ the regret of online kernel selection in a continuous kernel space. In contrast to traditional online kernel learning (Lu et al. 2016; Zhang and Liao 2019), the hypotheses generated by online kernel selection may lie in different reproducing kernel Hilbert spaces (RKHSs). Existing regret bounds of online kernel selection do not hold for a continuous kernel space due to the infinite number of candidate kernels. In the following section, we focus on the regret analyses of online kernel selection in a continuous kernel space under different learning frameworks.

Online Kernel Selection Categories in Continuous Kernel Space

In this section, we define the hypothesis sketch sequence for online kernel selection. To represent different learning frameworks of online kernel selection, we propose two categories of online kernel selection in a continuous kernel space with the hypothesis sketch sequence.

Hypothesis Sketch Sequence

In this subsection, we define a hypothesis sketch sequence for online kernel selection in a continuous kernel space. For a given continuous kernel space $\mathcal{K}_\Omega$, assuming that $\kappa_{\sigma_t} \in \mathcal{K}_\Omega$ is the kernel selected for prediction at round t, we first define *time-varying hypothesis sketch* at round t as follows:

$$f_{\sigma_t,t}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t)}(\cdot) \rangle, \quad \sigma_t \in \Omega,$$

where $\mathcal{V}_t = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^{|\mathcal{V}_t|} \subseteq \mathcal{X}$ is a buffer of size $|\mathcal{V}_t|$ ($|\mathcal{V}_t| \ll t$) at round t ,

$$\begin{aligned}\boldsymbol{\psi}_{\sigma_t}^{(t)}(\cdot) &= \left[\kappa_{\sigma_t}(\cdot, \tilde{\boldsymbol{x}}_1), \ldots, \kappa_{\sigma_t}(\cdot, \tilde{\boldsymbol{x}}_{|\mathcal{V}_t|})\right]^\mathsf{T}, \quad \tilde{\boldsymbol{x}}_i \in \mathcal{V}_t, \\ \boldsymbol{\omega}^{(t)} &= \left[\omega_1^{(t)}, \omega_2^{(t)}, \ldots, \omega_{|\mathcal{V}_t|}^{(t)}\right]^\mathsf{T} \in \mathbb{R}^{|\mathcal{V}_t|},\end{aligned}$$

are a *basis vector* and its corresponding *weight vector*, respectively. The hypothesis sketch $f_{\sigma_t,t}(\cdot)$ at round t can be seen as an approximation of the original hypothesis $h_{\sigma_t,t}(\cdot) = \sum_{i=1}^{t-1} \alpha_i^{(t)} \kappa_{\sigma_t}(\cdot, \boldsymbol{x}_i)$. After T rounds, we call $\{f_{\sigma_t,t}\}_{t=1}^T$ a *time-varying hypothesis sketch sequence* generated by online kernel selection. In contrast to the hypothesis sketch using a fixed kernel parameter per round, defined in (Zhang and Liao 2020), the time-varying hypothesis sketches we defined have time-varying kernel parameters at each round.

Two Online Kernel Selection Categories

In this subsection, we propose two categories of online kernel selection in a continuous kernel space with hypothesis sketch sequence. By considering all the possible orders of selection and training at each round, given a continuous kernel space $\mathcal{K}_\Omega$, we perform online kernel selection at round t in two different categories as follows: (a) Category 1 first

Category 1: Online Kernel Selection by Selection-Post-Training (OKS-SPT)

Require: the continuous kernel space $\mathcal{K}_\Omega$, initial kernel κ_{σ_1}

Initialize the weight vector $\boldsymbol{\omega}^{(1)} = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  Compute the hypothesis sketch $f_{\sigma_t,t}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t)}(\cdot) \rangle$
  Predict $\hat{y}_t = \mathrm{sgn}(f_{\sigma_t,t}(\boldsymbol{x}_t))$ for classification or $\hat{y}_t = f_{\sigma_t,t}(\boldsymbol{x}_t)$ for regression
  Maintain the buffer $\mathcal{V}_{t+1} = \textsc{BufferMaintenance}(\mathcal{V}_t, \boldsymbol{z}_t)$ and obtain $f_{\sigma_t,t}^{\mathrm{ma}} = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot) \rangle$
  Update the weight vector $\boldsymbol{\omega}^{(t+1)} = \textsc{WeightUpdating}(f_{\sigma_t,t}^{\mathrm{ma}}, \boldsymbol{z}_t)$ and obtain $f_{\sigma_t,t+1}^{\mathrm{ma}} = \langle \boldsymbol{\omega}^{(t+1)}, \boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot) \rangle$
  if the buffer changes **then**
    Select $\kappa_{\sigma_{t+1}} = \textsc{KernelSelection}(\kappa_{\sigma_t}, \boldsymbol{z}_t)$ from $\mathcal{K}_\Omega$ and obtain $f_{\sigma_{t+1},t+1}$
  else
    $\kappa_{\sigma_{t+1}} = \kappa_{\sigma_t}$
  end if
end for

maintains the buffer $\mathcal{V}_t$, obtains the following *maintained hypothesis sketch* with a new basis vector

$$f_{\sigma_t,t}^{\mathrm{ma}}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot) \rangle,$$
$$\boldsymbol{\psi}_{\sigma_t}^{(t+1)}(\cdot) = \left[\kappa_{\sigma_t}(\cdot, \tilde{\boldsymbol{x}}_1), \ldots, \kappa_{\sigma_t}(\cdot, \tilde{\boldsymbol{x}}_{|\mathcal{V}_{t+1}|})\right]^\intercal, \ \tilde{\boldsymbol{x}}_i \in \mathcal{V}_{t+1},$$

updates the weight vector of the maintained hypothesis sketch, and then selects a new RKHS associated with $\kappa_{\sigma_{t+1}}$ only when $\mathcal{V}_{t+1} \neq \mathcal{V}_t$, termed OKS-SPT; (b) Category 2 first selects the optimal kernel using a kernel selection criterion and the newly arrived instances, obtains $f_{\sigma_{t+1},t}^{\mathrm{ma}}$ by $\textsc{BufferMaintenance}$ in the new RKHS, and updates the weight vector, termed OKS-TPS. The main differences between the two categories are the order of kernel selection and hypothesis training at each round (see Figure 1) and the frequency of performing kernel selection, which result in completely different conditions for sublinear regret guarantees. We will specify the main steps of OKS-SPT and OKS-TPS, including $\textsc{BufferMaintenance}$, $\textsc{KernelSelection}$ and $\textsc{WeightUpdating}$.

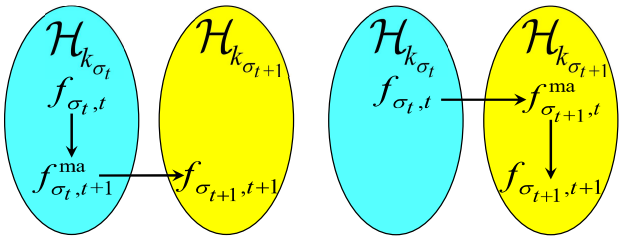

Figure 1: Comparison between OKS-SPT (left) and OKS-TPS (right) at round t, where OKS-SPT denotes the online kernel selection by selection-post-training and OKS-TPS denotes the online kernel selection by training-post-selection.

Category 2: Online Kernel Selection by Training-Post-Selection (OKS-TPS)

Require: the continuous kernel space $\mathcal{K}_\Omega$, initial kernel κ_{σ_1}

Initialize the weight vector $\boldsymbol{\omega}^{(1)} = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  Compute the hypothesis sketch $f_{\sigma_t,t}(\cdot) = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_t}^{(t)}(\cdot) \rangle$
  Predict $\hat{y}_t = \mathrm{sgn}(f_{\sigma_t,t}(\boldsymbol{x}_t))$ for classification or $\hat{y}_t = f_{\sigma_t,t}(\boldsymbol{x}_t)$ for regression
  Select a kernel $\kappa_{\sigma_{t+1}} = \textsc{KernelSelection}(\kappa_{\sigma_t}, \boldsymbol{z}_t)$ from $\mathcal{K}_\Omega$
  Maintain the buffer $\mathcal{V}_{t+1} = \textsc{BufferMaintenance}(\mathcal{V}_t, \boldsymbol{z}_t)$ and obtain $f_{\sigma_{t+1},t}^{\mathrm{ma}} = \langle \boldsymbol{\omega}^{(t)}, \boldsymbol{\psi}_{\sigma_{t+1}}^{(t+1)}(\cdot) \rangle$
  Update the weight vector $\boldsymbol{\omega}^{(t+1)} = \textsc{WeightUpdating}(f_{\sigma_{t+1},t}^{\mathrm{ma}}, \boldsymbol{z}_t)$ and obtain $f_{\sigma_{t+1},t+1}$
end for

Regret Analyses for the Two Categories

In this section, we formulate a surrogate hypothesis space for regret analyses in a continuous kernel space, analyze the regrets of the two online kernel selection categories in the surrogate hypothesis space, and compare our regret guarantees and computational complexities with those of the existing online kernel selection approaches. The detailed proofs of the theorems can be found in the supplementary material.

Surrogate Hypothesis Space

Since online kernel selection dynamically selects the optimal kernel at each round, the hypothesis sketches generated by online kernel selection may lie in different RKHSs. This poses new challenges of bounding the regret for online kernel selection due to the unknown bounds of norms and inner products of kernel functions in varying RKHSs. To address these issues, we construct a surrogate hypothesis space containing all the candidate kernels, and formulate the regret in a surrogate hypothesis space for online kernel selection. In the following section, we represent the parameter interval of candidate kernels by $\Omega = [\sigma_{\min}, \sigma_{\max}]$. The *surrogate hypothesis space* $\widehat{\mathcal{H}}$ corresponding to a continuous kernel space $\mathcal{K}_\Omega$ is the union of all the candidate RKHSs, i.e., $\widehat{\mathcal{H}} = \bigcup_{\kappa_\sigma \in \mathcal{K}_\Omega} \mathcal{H}_{\kappa_\sigma}$. Let $\sigma_t \in \Omega$ be the optimal kernel parameter used at round t, $\{f_{\sigma_t,t}\}_{t=1}^T \subseteq \widehat{\mathcal{H}}$ be a hypothesis sketch sequence generated by the proposed online kernel selection categories. We define the *regret in the surrogate hypothesis space* $\widehat{\mathcal{H}}$ for online kernel selection with $\{f_{\sigma_t,t}\}_{t=1}^T \subseteq \widehat{\mathcal{H}}$ in the form

$$\begin{aligned} &\widehat{\mathrm{Reg}}_T(\{f_{\sigma_t,t}\}_{t=1}^T, f^*) \\ &= \sum_{t=1}^T \left[\ell\left(f_{\sigma_t,t}(\boldsymbol{x}_t), y_t\right) - \ell\left(f^*(\boldsymbol{x}_t), y_t\right)\right], \end{aligned} \tag{1}$$

where the competing hypothesis $f^* \in \widehat{\mathcal{H}}$ is defined as $f^* = \arg\min_{f \in \widehat{\mathcal{H}}} \sum_{t=1}^T \ell\left(f(\boldsymbol{x}_t), y_t\right)$.

We give an example of a surrogate hypothesis space induced by the Gaussian kernel $\kappa_\sigma(\boldsymbol{x}_1, \boldsymbol{x}_2) =$

$\exp\left(-\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2/2\sigma^2\right)$, $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}$, where $\sigma>0$ is the kernel parameter of Gaussian kernel κ_σ. In contrast to Theorem 2 in (Zhang and Liao 2018), we analyze the surrogate hypothesis space in a more general framework, in which we construct the hypothesis space using a varying kernel parameter and bound both the norms and inner products, as shown in Theorem 1.

Theorem 1. *Let $\xi\in(0,1]$, $\mathcal{K}_\Omega=\{\kappa_\sigma\mid\sigma\in\Omega=[\sigma_{\min},\sigma_{\max}]\}$ be a continuous kernel space containing Gaussian kernels. Then, the RKHS $\widehat{\mathcal{H}}$ induced by the Gaussian kernel $\hat{\kappa}$ with kernel parameter $\hat{\sigma}=\sqrt{\xi}\sigma_{\min}$ is the surrogate hypothesis space containing all the candidate kernels in $\mathcal{K}_\Omega$, and for $\kappa_\sigma\in\mathcal{K}_\Omega$, $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}$, the following bounds hold*

$$\begin{aligned}\kappa_\sigma(\boldsymbol{x}_1,\boldsymbol{x}_2)\le\langle\kappa_\sigma(\cdot,\boldsymbol{x}_1),\kappa_\sigma(\cdot,\boldsymbol{x}_2)\rangle_{\widehat{\mathcal{H}}}\le\tau,\\ \|\kappa_\sigma(\cdot,\boldsymbol{x}_1)\|^2_{\widehat{\mathcal{H}}}=\tau,\end{aligned}\tag{2}$$

where $\tau=\left[2\left(\sigma_{\min}/\sigma\right)^2\xi-\left(\sigma_{\min}/\sigma\right)^4\xi^2\right]^{-\frac{d}{2}}$.

Remark 1. *From* (2)*, for all $\boldsymbol{x}\in\mathcal{X}$, we have*

$$\|\kappa_\sigma(\cdot,\boldsymbol{x})\|_{\mathcal{H}_{\kappa_\sigma}}\le\|\kappa_\sigma(\cdot,\boldsymbol{x})\|_{\widehat{\mathcal{H}}}\le\sqrt{\tau}\,\|\kappa_\sigma(\cdot,\boldsymbol{x})\|_{\mathcal{H}_{\kappa_\sigma}}.$$

Thus, the norms $\|\cdot\|_{\widehat{\mathcal{H}}}$ and $\|\cdot\|_{\mathcal{H}_{\kappa_\sigma}}$ are equivalent for Gaussian kernels $\kappa_\sigma\in\mathcal{K}_\Omega$.

Regret Bounds of Online Kernel Selection

In this subsection, we bound the regrets of different categories of online kernel selection in the surrogate hypothesis space. As implementations of the proposed online kernel selection categories, we select the optimal kernel at each round using a learning kernel approach. More specifically, we perform WEIGHTUPDATING using Kernelized Online Gradient Descent (KOGD) (Kivinen, Smola, and Williamson 2001) with a specific compensation strategy, and implement KERNELSELECTION by Online Gradient Descent (OGD) (Shalev-Shwartz 2011) over the instantaneous loss once the loss is convex with respect to the kernel parameter.

Regret Bound for OKS-SPT (Category 1) In OKS-SPT, at round t, we maintain the buffer and implement online kernel selection via learning kernel in the following three steps:

1. BUFFERMAINTENANCE: If some condition holds, insert the example $\boldsymbol{x}_t$ into the buffer $\mathcal{V}_t$ at round t without deletion.
2. WEIGHTUPDATING: If inserting $\boldsymbol{x}_t$ into $\mathcal{V}_t$, update the weight vector
$$f_{\sigma_t,t+1}(\cdot)=f^{\mathrm{ma}}_{\sigma_t,t}(\cdot)-\eta_f\nabla_{f^{\mathrm{ma}}_{\sigma_t,t}}\ell(f^{\mathrm{ma}}_{\sigma_t,t}(\boldsymbol{x}_t),y_t);$$
otherwise, $f_{\sigma_t,t+1}(\cdot)=f^{\mathrm{ma}}_{\sigma_t,t}(\cdot)+\delta_t(\cdot)$, where $\delta_t(\cdot)$ is a compensation to the weight vector at round t and $\eta_f>0$ is the stepsize of KOGD.
3. KERNELSELECTION: If the buffer changes, i.e., $\mathcal{V}_{t+1}\neq\mathcal{V}_t$, select the kernel using OGD
$$\sigma_{t+1}=\sigma_t-\eta_\sigma\nabla_{\sigma_t}\ell(f_{\sigma_t,t+1}(\boldsymbol{x}_t),y_t),$$
where $\nabla_{\sigma_t}\ell(f_{\sigma_t,t+1}(\boldsymbol{x}_t),y_t)$ is the gradient or a subgradient of $\ell(f_{\sigma_t,t+1}(\boldsymbol{x}_t),y_t)$ at σ_t and $\eta_\sigma>0$ is the stepsize of OGD.

We first analyze the instantaneous regrets for online kernel selection via learning kernel in the surrogate hypothesis space $\widehat{\mathcal{H}}$. For convenience, we define three types of best-in-hindsight hypotheses in different RKHSs as follows:

Best-in-hindsight hypothesis in $\widehat{\mathcal{H}}$: The best-in-hindsight hypothesis $\bar{f}^*\in\widehat{\mathcal{H}}$ and its corresponding kernel parameter $\bar{\sigma}^*$ are defined by

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

where $\bar{f}^*(\cdot)=\langle\bar{\boldsymbol{\omega}}^*,\boldsymbol{\psi}^*_{\bar{\sigma}^*}(\cdot)\rangle$ and

$$\boldsymbol{\psi}^*_{\bar{\sigma}^*}(\cdot)=[\kappa_{\bar{\sigma}^*}(\cdot,\boldsymbol{x}_1),\kappa_{\bar{\sigma}^*}(\cdot,\boldsymbol{x}_2),\ldots,\kappa_{\bar{\sigma}^*}(\cdot,\boldsymbol{x}_T)]^\intercal.$$

Best-in-hindsight hypothesis in $\mathcal{H}_{\kappa_{\sigma_t}}$: The best-in-hindsight hypothesis in $\mathcal{H}_{\kappa_{\sigma_t}}$ is denoted by

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

which can be expressed as $f^*_{\sigma_t}(\cdot)=\langle\boldsymbol{\omega}^*_{\sigma_t},\boldsymbol{\psi}^*_{\sigma_t}(\cdot)\rangle$.

Modified best-in-hindsight hypothesis in $\mathcal{H}_{\kappa_{\sigma_t}}$: The hypothesis $\bar{f}^*_{\sigma_t}(\cdot)=\langle\bar{\boldsymbol{\omega}}^*,\boldsymbol{\psi}^*_{\sigma_t}(\cdot)\rangle\in\mathcal{H}_{\kappa_{\sigma_t}}$, which uses the same kernel and basis vector as in $f^*_{\sigma_t}$ but the weight vector of $\bar{f}^*$.

For regret analyses, we define the *instantaneous regret* of f_a against f_b at round t as follows:

$$\mathrm{Reg}_t(f_a,f_b)=\ell\left(f_a(\boldsymbol{x}_t),y_t\right)-\ell\left(f_b(\boldsymbol{x}_t),y_t\right).$$

Then, we transform the regret in (1) for online kernel selection into

$$\widehat{\mathrm{Reg}}_T(\{f_{\sigma_t,t}\}_{t=1}^T,\bar{f}^*)=\sum_{t=1}^{T}\mathrm{Reg}_t(f_{\sigma_t,t},\bar{f}^*),$$

and split $\mathrm{Reg}_t(f_{\sigma_t,t},\bar{f}^*)$ into three instantaneous regrets

$$\begin{aligned}&\mathrm{Reg}_t(f_{\sigma_t,t},\bar{f}^*)\\&=\underbrace{\mathrm{Reg}_t(f_{\sigma_t,t},f^*_{\sigma_t})}_{\text{Optimization}}+\underbrace{\mathrm{Reg}_t(f^*_{\sigma_t},\bar{f}^*_{\sigma_t})}_{\text{Estimation}}+\underbrace{\mathrm{Reg}_t(\bar{f}^*_{\sigma_t},\bar{f}^*)}_{\text{Approximation}}.\end{aligned}\tag{3}$$

These three instantaneous regrets measure the performances of optimization, estimation and approximation in OKS-SPT, respectively, and the detailed interpretations of the three instantaneous regrets are given in the supplementary material.

Then, we define the *hypothesis sketch degradation* Δ_t at round t for OKS-SPT as follows:

$$\Delta_t=f_{\sigma_t,t}(\cdot)-\eta_f\nabla_{f_{\sigma_t,t}}\ell(f_{\sigma_t,t}(\boldsymbol{x}_t),y_t)-f_{\sigma_t,t+1}(\cdot),$$

and denote the *gradient error* and the *average gradient error* with respect to the hypothesis sketch by

$$\left\|E^{(t)}_{\mathrm{h}}\right\|_{\widehat{\mathcal{H}}}=\|\Delta_t/\eta_f\|_{\mathcal{H}_{\widehat{\mathcal{H}}}}\ \text{ and }\ \overline{E}_{\mathrm{h}}=\sum_{t=1}^{T}\left\|E^{(t)}_{\mathrm{h}}\right\|_{\mathcal{H}_{\widehat{\mathcal{H}}}}/T.$$

In order to obtain regret guarantees, we make the following assumption on OKS-SPT.

Assumption 1. *Let $\mu \in [0, \max_{\kappa_\sigma \in \mathcal{K}_\Omega} \kappa_\sigma(\boldsymbol{x}, \boldsymbol{x})]$, $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{K}_{\sigma,\mathcal{V}_t} = (\kappa_\sigma(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j))$ be the kernel matrix with sizes $|\mathcal{V}_t| \times |\mathcal{V}_t|$ where $\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j \in \mathcal{V}_t$. OKS-SPT satisfies the three conditions:*

- BUFFERMAINTENANCE *inserts the example as $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{\boldsymbol{x}_t\}$ at round t if*[1]

$$\det \boldsymbol{K}_{\sigma_t,\mathcal{V}_t \cup \{\boldsymbol{x}_t\}} / \det \boldsymbol{K}_{\sigma_t,\mathcal{V}_t} > \mu. \tag{4}$$

- KERNELSELECTION *generates a sequence of kernel parameters $\{\sigma_t\}_{t=1}^T$ such that $\max_{t\in[T]} |\nabla_{\sigma_t} f_{\sigma_t,t+1}(\boldsymbol{x}_t)| \leq L$, and $\sigma_{\max} < \min_{\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}}_j \in \mathcal{V}, \tilde{\boldsymbol{x}}_j \neq \tilde{\boldsymbol{x}}_i} \|\tilde{\boldsymbol{x}}_j - \tilde{\boldsymbol{x}}_i\|/\sqrt{3}$ for Gaussian kernels, where $\mathcal{V} = \cup_{t\in[T]} \mathcal{V}_t$.*
- *When* (4) *does not hold at round t,* WEIGHTUPDATING *compensates the weight vector using $\delta_t(\cdot)$ to minimize $\|E_{\mathrm{h}}^{(t)}\|_{\mathcal{H}_{\kappa_{\sigma_t}}}$.*

Finally, we choose the RKHS $\widehat{\mathcal{H}}$ induced by the Gaussian kernel $\hat{\kappa}$ with kernel parameter $\hat{\sigma} = \sqrt{\xi}\sigma_{\min}$ ($\xi \in (0, 1]$) as the surrogate hypothesis space as in Theorem 1, bound the three instantaneous regrets in the surrogate hypothesis space, and prove the regret bound of the proposed OKS-SPT via learning kernel under Assumption 1.

Theorem 2. *Let $\ell(\cdot,\cdot)$ be the hinge loss function, $\mathcal{Y} = \{-1, 1\}$, $\mathcal{K}_\Omega$ be a continuous kernel space that contains Gaussian kernel functions, $\{f_{\sigma_t,t}\}_{t=1}^T \subseteq \widehat{\mathcal{H}}$ be the hypothesis sketch sequence generated by OKS-SPT satisfying Assumption 1. Define $\widehat{D}(f, g) = \max_{t\in[T]} |f(\boldsymbol{x}_t) - g(\boldsymbol{x}_t)|$, where $f, g \in \widehat{\mathcal{H}}$. Assume $C_{\max} = \max_{i,j\in[T]} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ and $R = \sup_{f\in\widehat{\mathcal{H}}} \|f\|_{\widehat{\mathcal{H}}}$, then there exists a constant $C > 0$ such that*

$$\widehat{\mathrm{Reg}}_T(\{f_{\sigma_t,t}\}_{t=1}^T, \bar{f}^*) \leq U_1 + U_2 + U_3,$$

where $U_2 = \widehat{D}(\bar{f}^*_{\sigma_t}, f^*_{\sigma_t})$,

$$U_1 = 2RC\sqrt{\mu}\,[T - O(\ln T)] + \frac{R^2}{2\eta_f} + \frac{\left(C\sqrt{\mu}+1\right)^2}{2}\eta_f T,$$

$$U_3 = 2C_{\max}\frac{\sigma_{\max}}{\sigma_{\min}^3}\widehat{D}(\bar{f}^*_{\sigma_t}, f_{\sigma_t,t+1}) + \frac{(\bar{\sigma}^*)^2}{2\eta_\sigma} + \frac{\left[\sigma_{\min}^{-3} C_{\max}\widehat{D}(\bar{f}^*_{\sigma_t}, f_{\sigma_t,t+1}) + L\right]^2}{2}\eta_\sigma T.$$

Remark 2. *Setting the values of the hypotheses to $O(1/\sqrt{T})$ is a common assumption (Zhao et al. 2012; Hu et al. 2015), since it has no influence on the prediction when multiplying the weights by a factor of order $O(1/\sqrt{T})$ of magnitude. Thus, we assume that $|f^*_{\sigma_t}(\boldsymbol{x}_t)|$, $|\bar{f}^*_{\sigma_t}(\boldsymbol{x}_t)|$ and $|f_{\sigma_t,t+1}(\boldsymbol{x}_t)|$ are of order $O(1/\sqrt{T})$ for $t \in [T]$, set $\eta_f, \eta_\sigma = O(1/\sqrt{T})$ and $\mu = O(1/T)$, and obtain a $O(\sqrt{T})$ regret bound for online kernel selection with OKS-SPT. This regret bound holds for a continuous kernel space and it is optimal for a convex objective function and OGD (Hazan 2016).*

[1]From the bordered matrix inverse formula, $\boldsymbol{K}_{\sigma_t,\mathcal{V}_t}$ is nonsingular at each round.

Regret Bound for OKS-TPS (Category 2) In OKS-TPS, we perform online kernel selection at round t in three steps:

1. KERNELSELECTION: Select the kernel using OGD
$$\sigma_{t+1} = \sigma_t - \eta_\sigma \nabla_{\sigma_t} \ell(f_{\sigma_t,t}(\boldsymbol{x}_t), y_t),$$
where $\eta_\sigma > 0$ is the stepsize of OGD.
2. BUFFERMAINTENANCE: Insert the new example as $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{\boldsymbol{x}_t\}$ if some condition holds otherwise set $\mathcal{V}_{t+1} = \mathcal{V}_t$, and obtain $f^{\mathrm{ma}}_{\sigma_{t+1},t}(\cdot)$.
3. WEIGHTUPDATING: If $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{\boldsymbol{x}_t\}$, update the weight vector with KOGD
$$f_{\sigma_{t+1},t+1}(\cdot) = f^{\mathrm{ma}}_{\sigma_{t+1},t}(\cdot) - \eta_f \nabla_{f^{\mathrm{ma}}_{\sigma_{t+1},t}} \ell(f^{\mathrm{ma}}_{\sigma_{t+1},t}(\boldsymbol{x}_t), y_t);$$
otherwise, $f_{\sigma_{t+1},t+1}(\cdot) = f^{\mathrm{ma}}_{\sigma_{t+1},t}(\cdot) + \theta_t(\cdot)$, where $\theta_t(\cdot)$ is a compensation to the weight vector at round t and $\eta_f > 0$ is the stepsize of KOGD.

We first decompose $\mathrm{Reg}_t(f_{\sigma_t,t}, \bar{f}^*)$ for OKS-TPS into two instantaneous regrets as follows:

$$\mathrm{Reg}_t(f_{\sigma_t,t}, \bar{f}^*) = \underbrace{\mathrm{Reg}_t(f_{\sigma_t,t}, f_{\bar{\sigma}^*,t})}_{\text{Approximation}} + \underbrace{\mathrm{Reg}_t(f_{\bar{\sigma}^*,t}, \bar{f}^*)}_{\text{Optimization, Estimation}}.$$

In contrast to the three instantaneous regrets (3) in OKS-SPT, the two instantaneous regrets in OKS-TPS have different interpretations, given in the supplementary material.

Then, we define the *hypothesis sketch degradation* Λ_t at round t for OKS-TPS in a different form as follows:

$$\Lambda_t = f_{\sigma_{t+1},t}(\cdot) - \eta_f \nabla_{f_{\sigma_{t+1},t}} \ell(f_{\sigma_{t+1},t}(\boldsymbol{x}_t), y_t) - f_{\sigma_{t+1},t+1}(\cdot).$$

The corresponding *gradient error* and *average gradient error* are

$$\left\|F_{\mathrm{h}}^{(t)}\right\|_{\widehat{\mathcal{H}}} = \|\Lambda_t/\eta_f\|_{\mathcal{H}_{\widehat{\kappa}}} \quad \text{and} \quad \overline{F}_{\mathrm{h}} = \sum_{t\in[T]} \left\|F_{\mathrm{h}}^{(t)}\right\|_{\mathcal{H}_{\widehat{\kappa}}}/T,$$

respectively. For a sublinear regret guarantee, OKS-TPS requires different assumptions from OKS-SPT as follows.

Assumption 2. *Let $\nu \in [0, \max_{\kappa_\sigma \in \mathcal{K}_\Omega} \kappa_\sigma(\boldsymbol{x}, \boldsymbol{x})]$, $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{K}_{\sigma_t,\mathcal{V}_t} = (\kappa_{\sigma_t}(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j))$ be the kernel matrix with size $|\mathcal{V}_t| \times |\mathcal{V}_t|$ where $\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j \in \mathcal{V}_t$. OKS-TPS satisfies the three conditions:*

- BUFFERMAINTENANCE *inserts the example as $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \{\boldsymbol{x}_t\}$ at round t if*[2]

$$\det \boldsymbol{K}_{\sigma_{t+1},\mathcal{V}_t \cup \{\boldsymbol{x}_t\}} / \det \boldsymbol{K}_{\sigma_{t+1},\mathcal{V}_t} > \nu. \tag{5}$$

- KERNELSELECTION *generates a sequence of kernel parameters $\{\sigma_t\}_{t=1}^T$ such that $\max_{t\in[T]} |\nabla_{\sigma_t} f_{\sigma_t,t}(\boldsymbol{x}_t)| \leq M$, and $\sigma_t < \min_{\tilde{\boldsymbol{x}}_i\in\mathcal{V}_t} \|\boldsymbol{x}_t - \tilde{\boldsymbol{x}}_i\|/\sqrt{3}$ for Gaussian kernels and $t \in [T]$.*
- *When* (5) *does not hold at round t,* WEIGHTUPDATING *compensates the weight vector using $\theta_t(\cdot)$ to minimize $\|F_{\mathrm{h}}^{(t)}\|_{\mathcal{H}_{\kappa_{\sigma_{t+1}}}}$. For hinge loss, this condition is equivalent to $\theta_t(\cdot) = \eta_f y_t \langle \boldsymbol{\beta}_t^{\sigma_{t+1}}, \boldsymbol{\psi}^{(t)}_{\sigma_{t+1}}(\cdot)\rangle$, where $\boldsymbol{\beta}_t^{\sigma_{t+1}} = \left(\boldsymbol{K}_{\sigma_{t+1},\mathcal{V}_t}\right)^\dagger \boldsymbol{\psi}^{(t)}_{\sigma_{t+1}}(\boldsymbol{x}_t)$.*

[2]From the bordered matrix inverse formula, $\boldsymbol{K}_{\sigma_{t+1},\mathcal{V}_t}$ is nonsingular at each round.

Approach	Computational complexities				Regret guarantees	
	Time (round t)	#Updates	Time (overall)	Space	Candidate	Regret bound
OKS (Yang et al. 2012)	$O(N+t)$	T	$O(T^2+NT)$	$O(T)$	Finite	$O\left(\sqrt{N(\ln N)T}\right)$
MS-FTPL (Foster et al. 2017)	$O\left(Nt\right)$	T	$O\left(NT^2\right)$	$O(NT)$	Finite	$O\left(\sqrt{NT\ln T}\right)$
OKL-GD (Chen et al. 2016)	$O(t)$	T	$O(T^2)$	$O(T)$	Continuous	–
RRF (Nguyen et al. 2017)	$O(D)$	T	$O(DT)$	$O(D)$	Continuous	–
OKS-SPT	$O\left((\ln t)^2\right)$	$\ln(T)$	$O((\ln T)^2T)$	$O((\ln T)^2)$	Continuous	$O\left(\sqrt{T}\right)$
OKS-TPS	$O\left((\ln t)^2\right)$	T	$O((\ln T)^2T)$	$O((\ln T)^2)$	Continuous	$O\left(\sqrt{T}\right)$

Table 1: Comparisons between the proposed categories with learning kernel and the existing approaches for online kernel selection, where MS-FTPL uses a uniform prior distribution. (T: the number of rounds; N: the number of candidate kernels, $N<\infty$; D: the dimension of random feature space; #Updates: the number of updates for optimal kernels; Time: time complexity; Space: space complexity; Candidate: the types of the set of candidate kernels; Finite: the finite kernel set; Continuous: the continuous kernel space; "–": not available).

We finally give the bounds of the two instantaneous regrets in the surrogate hypothesis space, and derive the regret bound of OKS-TPS under Assumption 2.

Theorem 3. *Let $\ell(\cdot,\cdot)$ be the hinge loss function, $\mathcal{Y}=\{-1,1\}$, $\mathcal{K}_\Omega$ be a continuous kernel space that contains Gaussian kernel functions, $\{f_{\sigma_t,t}\}_{t=1}^T\subseteq\widehat{\mathcal{H}}$ be the hypothesis sketch sequence generated by OKS-TPS satisfying Assumption 2. Assume $R=\sup_{f\in\widehat{\mathcal{H}}}\|f\|_{\widehat{\mathcal{H}}}$, then*

$$\widehat{\mathrm{Reg}}_T(\{f_{\sigma_t,t}\}_{t=1}^T,\bar{f}^*)\leq J_1+J_2,$$

where $J_1=(\bar{\sigma}^*)^2/2\eta_\sigma+M^2\eta_\sigma T/2$ and

$$J_2=2R\sum_{t=1}^T D_t^*+2RTO(\sqrt{\nu})+\frac{R^2}{2\eta_f}+\frac{\left[\max_{t\in[T]}D_t^*+O(\sqrt{\nu})+1\right]^2}{2}\eta_f T,$$

where $D_t^*=0$ if $\mathcal{V}_{t+1}=\mathcal{V}_t\cup\{\boldsymbol{x}_t\}$ and otherwise

$$D_t^*=\left(\frac{\det\boldsymbol{K}_{\bar{\sigma}^*,\mathcal{V}_t\cup\{\boldsymbol{x}_t\}}}{\det\boldsymbol{K}_{\bar{\sigma}^*,\mathcal{V}_t}}+\left\langle\boldsymbol{\psi}_{\bar{\sigma}^*}^{(t)}(\boldsymbol{x}_t),\,\boldsymbol{\beta}_t^{\bar{\sigma}^*}-\boldsymbol{\beta}_t^{\sigma_{t+1}}\right\rangle\right)^{\frac{1}{2}}.$$

Remark 3. *In contrast to OKS-SPT, the optimal regret bound of OKS-TPS does not need the assumptions for the values of the hypotheses. Setting $\eta_f,\eta_\sigma=O(1/\sqrt{T})$ and $\nu=O(1/T)$, if the following conditions hold when* (5) *holds at round t*

$$\begin{aligned}\det\boldsymbol{K}_{\bar{\sigma}^*,\mathcal{V}_t\cup\{\boldsymbol{x}_t\}}/\det\boldsymbol{K}_{\bar{\sigma}^*,\mathcal{V}_t}&=O(\nu),\\ \|\boldsymbol{\beta}_t^{\bar{\sigma}^*}-\boldsymbol{\beta}_t^{\sigma_{t+1}}\|&=O(\nu),\end{aligned}\tag{6}$$

OKS-TPS enjoys a $O(\sqrt{T})$ regret bound in a continuous kernel space, which is optimal for a convex objective function and OGD (Hazan 2016). (6) *measures the quality of the continuous kernel space, which can be verified only using candidate kernels and the examples without labels.*

Comparisons with Existing Theoretical Results

In this subsection, we summarize the comparable theoretical results for the online kernel selection, including computational complexities[3] and regret guarantees .

For our two categories of online kernel selection, the running time is dominated by the computing the determinants in (4) and (5). From the proofs of Theorem 2 and Theorem 3, we can obtain that the number of the examples that satisfy (4) is $O(\ln t)$ after t rounds. In practical implementations, we use $\sigma_{\min}$ instead of σ_t or σ_{t+1} in the conditions (4), (5) and the compensations $\delta_t(\cdot)$, $\theta_t(\cdot)$. Then the determinants can be computed using rank-one Cholesky updates (Golub and Van Loan 2012), resulting in a $O\left((\ln t)^2\right)$ time complexity at round t and a polylogarithmic space complexity for both categories of online kernel selection. Besides, since OKS-SPT only needs a logarithmic number of updates for optimal kernels, it has a $O((\ln T)^3)$ overall time complexity for kernel selection which is more efficient than OKS-TPS.

In regret analysis, we focus on the analyses and comparisons of worse-case regrets which are conceptually stronger than expected regret. The reason is that the expected regrets ignore the variance information. OKS-SIL in (Zhang and Liao 2020) can be seen as a special case of our Category 1, but OKS-SIL does not satisfy our condition (4) for sublinear worse-case regrets of Category 1, and only enjoys a weaker expected regret bound. Thus, we obtain stronger theoretical guarantees for online kernel selection in continuous kernel space. Table 1 summarizes the theoretical results of our two categories of online kernel selection and the existing approaches, from which we can observe the following results: (a) in contrast to the existing linear time complexity with respect to the number of candidate kernels, the time complexity of the proposed two categories is independent of the cardinality of the continuous kernel space, which is effective for a continuous kernel space; (b) the proposed two categories reduce the linear space complexity to a logarithmic space complexity with respect to the number of rounds, and have a polylogarithmic time complexity at each round with respect to the current number of rounds compared with the existing linear time complexity[4]; (c) unlike the existing

[3] We omit the dimension of input data in the computational complexities.

[4] For online kernel learning using random features, a dimension

Algorithm	german		spambase		mushrooms	
	Mistake rate (%)	Time (s)	Mistake rate (%)	Time (s)	Mistake rate (%)	Time (s)
OKL-GD	34.960 ± 1.518	**0.194**	36.031 ± 0.421	6.700	4.226 ± 0.910	37.230
OKS	42.320 ± 1.307	0.226	34.355 ± 0.372	4.083	9.441 ± 0.282	9.787
RRF	31.140 ± 0.114	0.372	44.961 ± 0.820	4.767	16.166 ± 0.964	21.750
OKS-SPT	29.920 ± 0.286	0.244	$\mathbf{28.436 \pm 0.213}$	**2.533**	6.585 ± 0.246	**4.240**
OKS-TPS	$\mathbf{29.760 \pm 0.270}$	0.296	28.450 ± 0.188	2.590	$\mathbf{3.139 \pm 0.481}$	6.910
Algorithm	a9a		w7a		ijcnn1	
	Mistake rate (%)	Time (s)	Mistake rate (%)	Time (s)	Mistake rate (%)	Time (s)
OKL-GD	23.936 ± 0.008	321.525	2.975 ± 0.062	857.268	9.575 ± 0.012	134.880
OKS	23.617 ± 0.127	1053.420	7.637 ± 0.024	943.855	9.578 ± 0.184	618.520
RRF	23.931 ± 0.001	152.265	2.978 ± 0.004	674.735	9.574 ± 0.001	39.290
OKS-SPT	$\mathbf{20.368 \pm 0.659}$	**39.360**	2.675 ± 0.023	**94.530**	9.478 ± 0.003	**33.860**
OKS-TPS	22.379 ± 0.192	48.815	$\mathbf{2.631 \pm 0.010}$	96.395	$\mathbf{9.440 \pm 0.002}$	35.165

Table 2: Performances of OKL-GD, OKS, RRF and the proposed OKS-SPT, OKS-TPS for online classification w.r.t. the mistake rate $= \sum_{t=1}^{T} I(y_t f_{\sigma_t,t}(\boldsymbol{x}_t) < 0)/T \times 100$ and the running time.

approaches for a continuous kernel space lacking sublinear regrets, the proposed two categories enjoy sublinear regrets in a continuous kernel space. Besides, although OKS-SPT requires more conditions for a sublinear regret guarantee than OKS-TPS, it is more efficient than OKS-TPS due to its lower overall time complexity for kernel selection.

Empirical Studies

This section empirically evaluates the performances of different categories of online kernel selection, verifying the correctness of the theoretical results. We merged the training set and testing set into one dataset for each benchmark dataset[5]. We performed the experiments over 20 different random permutations of the datasets, which were implemented in R 3.3.2 on a machine with 4-core Intel Core i7 3.60 GHz CPU and 16GB memory. We compared the proposed categories of online kernel selection with the following state-of-the-art online kernel selection algorithms: Online Kernel Learning with Gradient Descent[6] (OKL-GD) (Chen et al. 2016), Online Kernel Selection (OKS) (Yang et al. 2012), Reparameterized Random Feature (RRF) (Nguyen et al. 2017).

We adopted Gaussian kernels as the candidate kernels, including a finite kernel set $\{2^{-(i+1)/2}, i = [-14 : +2 : 14]\}$ for OKS and a continuous kernel space $\Omega = [2^{-15/2}, 2^{13/2}]$ for OKL-GD, RRF and our categories. For all the algorithms, we used the hinge loss functions, tuned the stepsizes of online gradient descent in a range of $10^{[-5:+1:0]}$, and selected the initial kernel σ_1 in $\{2^{-(i+1)/2}, i = [-14 : +1 : -10]\}$ uniform randomly, since small σ_1 may lead to the vanishing of the gradients. In our categories, we set $\mu = 0.1$ and used budgeted versions of the proposed categories that stop updating the buffer under a fixed budget $B = 200$. Besides, we set the dimension of random features $D = 400$ in RRF, and set the smoothing parameter and stepsize of OKS as in (Yang et al. 2012).

Table 2 lists the experimental results of the mistake rate and the running time for online classification on benchmark datasets. We summarize the observations as follows: (a) the proposed categories are more efficient on large datasets and more accurate on all the datasets than the other online kernel selection algorithms, which conforms to the theoretical results in Table 1; (b) OKS-SPT is more efficient than OKS-TPS on all the datasets. The reason is that OKS-SPT requires only polylogarithmic overall time complexity for kernel selection compared with the quasilinear overall time complexity of OKS-TPS for kernel selection; (c) OKS-TPS performs better than OKS-SPT in terms of the mistake rates on most datasets. This is because the sublinear regret bound of OKS-TPS requires less conditions to hold than OKS-SPT. which is analyzed in theoretical results.

Conclusion

Regret analysis for online kernel selection in a continuous kernel space is a brand-new and complex problem. In this paper, we divide online kernel selection in a continuous kernel space into two categories according to the order of selection and training at each round. We give the conditions that guarantee the optimal regret bounds for the two categories in continuous kernel spaces, and demonstrate that the two categories via the proposed kernel selection algorithms have polylogarithmic computational complexities at each round with respect to the current number of rounds. The theoretical results establish a solid foundation for the regret analytics of online model selection and online learning under different learning frameworks. Future work will extend our regret analyses to decision problem under limited feedbacks.

of random feature space of order $D = O(T)$ is required for a sublinear regret bound (Lu et al. 2016).

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

[6]For OKL-GD, we restricted the number of the support vectors to 2000, preventing the curse of kernelization.

Acknowledgments

This work was funded by the National Key R&D Program of China (2019YFE0198200), the National Natural Science Foundation of China (No. 62006234, No. 61872338, No. 61832017, No. 62076234), Beijing Academy of Artificial Intelligence (BAAI2019ZD0305), the Tencent WeChat Rhino-Bird Focused Research Program, China Association for Science and Technology "Youth Project of High-End Technology Innovation and Think Tank", and Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098.

References

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge University Press.

Chen, B.; Liang, J.; Zheng, N.; and Príncipe, J. C. 2016. Kernel least mean square with adaptive kernel size. *Neurocomputing* 191: 95–106.

Diethe, T.; and Girolami, M. 2013. Online learning with (multiple) kernels: A review. *Neural Computation* 25(3): 567–625.

Ding, L.; Liu, Y.; Liao, S.; Li, Y.; Yang, P.; Pan, Y.; Huang, C.; Shao, L.; and Gao, X. 2019. Approximate kernel selection with strong approximate consistency. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 3462–3469.

Foster, D. J.; Kale, S.; Mohri, M.; and Sridharan, K. 2017. Parameter-Free Online Learning via Model Selection. In *Advances in Neural Information Processing Systems 30*, 6022–6032.

Golub, G. H.; and Van Loan, C. F. 2012. *Matrix computations*. Johns Hopkins University Press.

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

Hu, J.; Yang, H.; King, I.; Lyu, M. R.; and So, A. M. 2015. Kernelized online imbalanced learning with fixed budgets. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2666–2672.

Jin, R.; Hoi, S. C.; and Yang, T. 2010. Online multiple kernel learning: Algorithms and mistake bounds. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory*, 390–404.

Kivinen, J.; Smola, A. J.; and Williamson, R. C. 2001. Online learning with kernels. In *Advances in Neural Information Processing Systems 14*, 785–792.

Liu, Y.; Liao, S.; Jiang, S.; Ding, L.; Lin, H.; and Wang, W. 2020. Fast cross-validation for kernel-based algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(5): 1083–1096.

Lu, J.; Hoi, S. C.; Wang, J.; Zhao, P.; and Liu, Z. 2016. Large scale online kernel learning. *Journal of Machine Learning Research* 17: 1613–1655.

Muthukumar, V.; Ray, M.; Sahai, A.; and Bartlett, P. L. 2019. Best of many worlds: Robust model selection for online supervised learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 3177–3186.

Nguyen, K. 2017. Nonparametric online machine learning with kernels. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 5197–5198.

Nguyen, T. D.; Le, T.; Bui, H.; and Phung, D. Q. 2017. Large-scale online kernel learning with random feature reparameterization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2543–2549.

Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 1177–1184.

Rakhlin, A.; Shamir, O.; and Sridharan, K. 2012. Relax and randomize: From value to algorithms. *Advances in Neural Information Processing Systems 25* 3: 2141–2149.

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

Singh, A.; and Príncipe, J. C. 2011. Information theoretic learning with adaptive kernels. *IEEE Transactions on Signal Processing* 91(2): 203–213.

Yang, T.; Mahdavi, M.; Jin, R.; Yi, J.; and Hoi, S. C. 2012. Online kernel selection: Algorithms and evaluations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 22–26.

Zhang, X.; and Liao, S. 2018. Online kernel selection via incremental sketched kernel alignment. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3118–3124.

Zhang, X.; and Liao, S. 2019. Incremental randomized sketching for online kernel learning. In *Proceedings of the 36th International Conference on Machine Learning*, 7394–7403.

Zhang, X.; and Liao, S. 2020. Hypothesis sketching for online kernel selection in continuous kernel space. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 2498–2504.

Zhang, X.; Liao, Y.; and Liao, S. 2019. A survey on online kernel selection for online kernel learning. *WIREs Data Mining and Knowledge Discovery* 9(2): e1295.

Zhao, P.; Wang, J.; Wu, P.; Jin, R.; and Hoi, S. C. 2012. Fast bounded online gradient descent algorithms for scalable kernel-Based online learning. In *Proceedings of the 29th International Conference on Machine Learning*, 169–176.