Supplementary Material for Incremental Randomized Sketching for Online Kernel Learning

Xiao Zhang [1] **Shizhong Liao** [1]

Abstract

In this supplementary material, we give the formulation of our incremental randomized sketching, the detailed proofs of the lemmas and theorems in the section "Theoretical Analysis", and more experimental results. Our main theoretical results include:

- The inner product preserving property (Lemma 2).
- The matrix product preserving property (Lemma 3).
- The low-rank approximation property (Theorem 1).
- The regret bound for online kernel learning (Theorem 2).

1. Formulation of Incremental Randomized Sketching

In this section, we construct the incremental randomized sketches of the kernel matrix, formulate the incremental maintenance for the incremental randomized sketches, and build the time-varying explicit feature mapping.

In the online setting, at round $t+1$, a new example $\boldsymbol{x}_{t+1}$ arrives and the kernel matrix $\boldsymbol{K}^{(t+1)}$ can be represented as a bordered matrix as follows:

$$\boldsymbol{K}^{(t+1)} = \begin{bmatrix} \boldsymbol{K}^{(t)} & \boldsymbol{\psi}^{(t+1)} \\ \boldsymbol{\psi}^{(t+1)\intercal} & \xi^{(t+1)} \end{bmatrix} \in \mathbb{R}^{(t+1)\times(t+1)},$$

where $\xi^{(t+1)} = \kappa(\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1})$ and

$$\boldsymbol{\psi}^{(t+1)} = [\kappa(\boldsymbol{x}_{t+1}, \boldsymbol{x}_1), \kappa(\boldsymbol{x}_{t+1}, \boldsymbol{x}_2), \ldots, \kappa(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t)]^\intercal.$$

First, we approximate the kernel matrix incrementally. Let $\boldsymbol{S}_{\mathrm{p}}^{(t+1)} \in \mathbb{R}^{(t+1)\times s_{\mathrm{p}}}$ be an SJLT and $\boldsymbol{S}_{\mathrm{m}}^{(t+1)} \in \mathbb{R}^{(t+1)\times s_{\mathrm{m}}}$ be a sub-sampling matrix. We use $\boldsymbol{S}_{\mathrm{p}}^{(t+1)}$ and $\boldsymbol{S}_{\mathrm{m}}^{(t+1)}$ for reducing the complexity of the problem (2) and the size of the approximate kernel matrix, respectively, and then formulate the *incremental randomized sketches* of $\boldsymbol{K}^{(t+1)}$ as follows:

$$\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)} = \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal}\boldsymbol{C}_{\mathrm{m}}^{(t+1)} \text{ and } \boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)} = \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal}\boldsymbol{C}_{\mathrm{p}}^{(t+1)},$$

where

$$\boldsymbol{C}_{\mathrm{p}}^{(t+1)} = \boldsymbol{K}^{(t+1)}\boldsymbol{S}_{\mathrm{p}}^{(t+1)} \text{ and } \boldsymbol{C}_{\mathrm{m}}^{(t+1)} = \boldsymbol{K}^{(t+1)}\boldsymbol{S}_{\mathrm{m}}^{(t+1)}.$$

Then $\boldsymbol{K}^{(t+1)}$ can be approximated by

$$\boldsymbol{K}_{\mathrm{sk}}^{(t+1)} = \boldsymbol{C}_{\mathrm{m}}^{(t+1)}\boldsymbol{F}_{\mathrm{sk}}^{(t+1)}\boldsymbol{C}_{\mathrm{m}}^{(t+1)\intercal} \approx \boldsymbol{K}^{(t+1)}, \quad (1)$$

where $\boldsymbol{F}_{\mathrm{sk}}^{(t+1)} \in \mathbb{R}^{s_{\mathrm{m}}\times s_{\mathrm{m}}}$ is obtained by solving the following sketched matrix approximation problem at round $t+1$

$$\begin{aligned} \boldsymbol{F}_{\mathrm{sk}}^{(t+1)} &= \arg\min_{\boldsymbol{F}} \left\| \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal}\boldsymbol{E}_{\boldsymbol{F}}^{(t+1)}\boldsymbol{S}_{\mathrm{p}}^{(t+1)} \right\|_{\mathrm{F}}^2 \\ &= \left(\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)}\right)^{\dagger} \boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)} \left(\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)\intercal}\right)^{\dagger}, \end{aligned} \quad (2)$$

where $\boldsymbol{E}_{\boldsymbol{F}}^{(t+1)}$ is the approximation error at round $t+1$

$$\boldsymbol{E}_{\boldsymbol{F}}^{(t+1)} = \boldsymbol{C}_{\mathrm{m}}^{(t+1)}\boldsymbol{F}\boldsymbol{C}_{\mathrm{m}}^{(t+1)\intercal} - \boldsymbol{K}^{(t+1)}.$$

We partition the sketch matrices into block matrices as

$$\boldsymbol{S}_{\mathrm{p}}^{(t+1)} = \left[\boldsymbol{S}_{\mathrm{p}}^{(t)\intercal}, \boldsymbol{s}_{\mathrm{p}}^{(t+1)}\right]^\intercal, \quad \boldsymbol{S}_{\mathrm{m}}^{(t+1)} = \left[\boldsymbol{S}_{\mathrm{m}}^{(t)\intercal}, \boldsymbol{s}_{\mathrm{m}}^{(t+1)}\right]^\intercal,$$

where $\boldsymbol{s}_{\mathrm{p}}^{(t+1)} \in \mathbb{R}^{s_{\mathrm{p}}}$ contains d nonzero entries that are determined by d different hash mappings in SJLT. Then the two incremental randomized sketches of the kernel matrix $\boldsymbol{K}^{(t+1)}$ can be computed incrementally by by rank-1 modifications as follows:

1) Sketch $\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)}$

[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. Correspondence to: Shizhong Liao <szliao@tju.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

The sketch $\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)}$ can be maintained as

$$\begin{aligned}
&\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)} \\
&= \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal} \boldsymbol{C}_{\mathrm{m}}^{(t+1)} \\
&= \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal} \boldsymbol{K}^{(t+1)} \boldsymbol{S}_{\mathrm{m}}^{(t+1)} \\
&= [\boldsymbol{S}_{\mathrm{p}}^{(t)\intercal}, \boldsymbol{s}_{\mathrm{p}}^{(t+1)}] \begin{bmatrix} \boldsymbol{K}^{(t)} & \boldsymbol{\psi}^{(t+1)} \\ \boldsymbol{\psi}^{(t+1)\intercal} & \xi^{(t+1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{\mathrm{m}}^{(t)} \\ \boldsymbol{s}_{\mathrm{m}}^{(t+1)\intercal} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{S}_{\mathrm{p}}^{(t)\intercal} \boldsymbol{K}^{(t)} + \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \boldsymbol{\psi}^{(t+1)\intercal} \\ \boldsymbol{S}_{\mathrm{p}}^{(t)\intercal} \boldsymbol{\psi}^{(t+1)} + \xi^{(t+1)} \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \end{bmatrix}^{\intercal} \begin{bmatrix} \boldsymbol{S}_{\mathrm{m}}^{(t)} \\ \boldsymbol{s}_{\mathrm{m}}^{(t+1)\intercal} \end{bmatrix} \\
&= \boldsymbol{S}_{\mathrm{p}}^{(t)\intercal} \boldsymbol{K}^{(t)} \boldsymbol{S}_{\mathrm{m}}^{(t)} + \boldsymbol{R}_{\mathrm{pm}}^{(t+1)} + \boldsymbol{R}_{\mathrm{mp}}^{(t+1)\intercal} + \boldsymbol{T}_{\mathrm{pm}}^{(t+1)} \\
&= \boldsymbol{\Phi}_{\mathrm{pm}}^{(t)} + \boldsymbol{R}_{\mathrm{pm}}^{(t+1)} + \boldsymbol{R}_{\mathrm{mp}}^{(t+1)\intercal} + \boldsymbol{T}_{\mathrm{pm}}^{(t+1)},
\end{aligned}$$

where the modifications are performed using the following three rank-1 matrices

$$\begin{aligned}
\boldsymbol{R}_{\mathrm{pm}}^{(t+1)} &= \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \boldsymbol{\psi}^{(t+1)\intercal} \boldsymbol{S}_{\mathrm{m}}^{(t)}, \\
\boldsymbol{R}_{\mathrm{mp}}^{(t+1)} &= \boldsymbol{s}_{\mathrm{m}}^{(t+1)} \boldsymbol{\psi}^{(t+1)\intercal} \boldsymbol{S}_{\mathrm{p}}^{(t)}, \\
\boldsymbol{T}_{\mathrm{pm}}^{(t+1)} &= \xi^{(t+1)} \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \boldsymbol{s}_{\mathrm{m}}^{(t+1)\intercal}.
\end{aligned}$$

2) Sketch $\boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)}$

For sketch $\boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)}$, we have

$$\begin{aligned}
&\boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)} \\
&= \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal} \boldsymbol{C}_{\mathrm{p}}^{(t+1)} \\
&= \boldsymbol{S}_{\mathrm{p}}^{(t+1)\intercal} \boldsymbol{K}^{(t+1)} \boldsymbol{S}_{\mathrm{p}}^{(t+1)} \\
&= [\boldsymbol{S}_{\mathrm{p}}^{(t)\intercal}, \boldsymbol{s}_{\mathrm{p}}^{(t+1)}] \begin{bmatrix} \boldsymbol{K}^{(t)} & \boldsymbol{\psi}^{(t+1)} \\ \boldsymbol{\psi}^{(t+1)\intercal} & \xi^{(t+1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{\mathrm{p}}^{(t)} \\ \boldsymbol{s}_{\mathrm{p}}^{(t+1)\intercal} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{S}_{\mathrm{p}}^{(t)\intercal} \boldsymbol{K}^{(t)} + \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \boldsymbol{\psi}^{(t+1)\intercal} \\ \boldsymbol{S}_{\mathrm{p}}^{(t)\intercal} \boldsymbol{\psi}^{(t+1)} + \xi^{(t+1)} \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \end{bmatrix}^{\intercal} \begin{bmatrix} \boldsymbol{S}_{\mathrm{p}}^{(t)} \\ \boldsymbol{s}_{\mathrm{p}}^{(t+1)\intercal} \end{bmatrix} \\
&= \boldsymbol{S}_{\mathrm{p}}^{(t)\intercal} \boldsymbol{K}^{(t)} \boldsymbol{S}_{\mathrm{p}}^{(t)} + \boldsymbol{R}_{\mathrm{pp}}^{(t+1)} + \boldsymbol{R}_{\mathrm{pp}}^{(t+1)\intercal} + \boldsymbol{T}_{\mathrm{pp}}^{(t+1)}, \\
&= \boldsymbol{\Phi}_{\mathrm{pp}}^{(t)} + \boldsymbol{R}_{\mathrm{pp}}^{(t+1)} + \boldsymbol{R}_{\mathrm{pp}}^{(t+1)\intercal} + \boldsymbol{T}_{\mathrm{pp}}^{(t+1)},
\end{aligned}$$

where the modifications are done by the following two rank-1 matrices

$$\begin{aligned}
\boldsymbol{R}_{\mathrm{pp}}^{(t+1)} &= \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \boldsymbol{\psi}^{(t+1)\intercal} \boldsymbol{S}_{\mathrm{p}}^{(t)}, \\
\boldsymbol{T}_{\mathrm{pp}}^{(t+1)} &= \xi^{(t+1)} \boldsymbol{s}_{\mathrm{p}}^{(t+1)} \boldsymbol{s}_{\mathrm{p}}^{(t+1)\intercal}.
\end{aligned}$$

Finally, we construct the time-varying explicit feature mapping using the incremental randomized sketches. We decompose $\boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)}$ via the rank-k singular value decomposition (SVD) as follows:

$$\boldsymbol{\Phi}_{\mathrm{pp}}^{(t+1)} \approx \boldsymbol{V}^{(t+1)} \boldsymbol{\Sigma}^{(t+1)} \boldsymbol{V}^{(t+1)\intercal},$$

where $\boldsymbol{V}^{(t+1)} \in \mathbb{R}^{s_{\mathrm{p}} \times k}$, $\boldsymbol{\Sigma}^{(t+1)} \in \mathbb{R}^{k \times k}$ and rank $k \leq s_{\mathrm{p}}$.

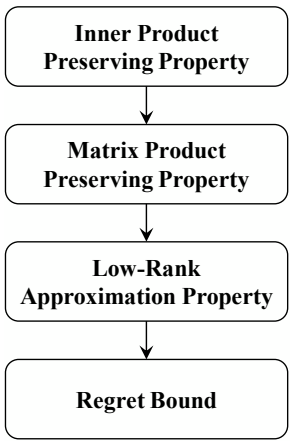


Figure 1. The dependence structure of our theoretical results.

Then $\boldsymbol{F}_{\mathrm{sk}}^{(t+1)}$ is approximated by

$$\boldsymbol{F}_{\mathrm{sk}}^{(t+1)} \approx \boldsymbol{Q}_{t+1} \boldsymbol{Q}_{t+1}^{\intercal},$$

where

$$\boldsymbol{Q}_{t+1} = \left(\boldsymbol{\Phi}_{\mathrm{pm}}^{(t+1)}\right)^{\dagger} \boldsymbol{V}^{(t+1)} \left(\boldsymbol{\Sigma}^{(t+1)}\right)^{\frac{1}{2}},$$

which yields the approximate kernel matrix from (1)

$$\boldsymbol{K}_{\mathrm{sk}}^{(t+1)} \approx \left(\boldsymbol{C}_{\mathrm{m}}^{(t+1)} \boldsymbol{Q}_{t+1}\right) \left(\boldsymbol{C}_{\mathrm{m}}^{(t+1)} \boldsymbol{Q}_{t+1}\right)^{\intercal}.$$

Thus, the kernel function value between the i-th example $\boldsymbol{x}_i$ and the j-th example $\boldsymbol{x}_j$ can be approximated by

$$\begin{aligned}
&\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&\approx \left(\left[\boldsymbol{C}_{\mathrm{m}}^{(t+1)}\right]_{i*} \boldsymbol{Q}_{t+1}\right) \left(\left[\boldsymbol{C}_{\mathrm{m}}^{(t+1)}\right]_{j*} \boldsymbol{Q}_{t+1}\right)^{\intercal},
\end{aligned}$$

and the explicit feature mapping can be updated at round $t+1$ by

$$\boldsymbol{\phi}_{t+2}(\cdot) = \left([\kappa(\cdot, \tilde{\boldsymbol{x}}_1), \ldots, \kappa(\cdot, \tilde{\boldsymbol{x}}_{s_{\mathrm{m}}})]\, \boldsymbol{Q}_{t+1}\right)^{\intercal},$$

where $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{s_{\mathrm{m}}}$ are the sampled examples obtained by $\boldsymbol{S}_{\mathrm{m}}^{(t+1)}$.

2. Detailed Proofs in Theoretical Analysis

Figure 1 describes the dependence structure of our theoretical results. For convenience, in this section, we denote $\boldsymbol{S}_{\mathrm{p}}^{(T)}$, $\boldsymbol{S}_{\mathrm{m}}^{(T)}$, $\boldsymbol{C}_{\mathrm{m}}^{(T)}$, $\boldsymbol{F}_{\mathrm{sk}}^{(T)}$ and $\boldsymbol{K}^{(T)}$ by $\boldsymbol{S}_{\mathrm{p}}$, $\boldsymbol{S}_{\mathrm{m}}$, $\boldsymbol{C}_{\mathrm{m}}$, $\boldsymbol{F}_{\mathrm{sk}}$ and $\boldsymbol{K}$, respectively. We first give some extra notations. Let $\boldsymbol{U}_{\mathrm{m}} \in \mathbb{R}^{T \times s_{\mathrm{m}}}$ be the first s_{m} left singular vectors of $\boldsymbol{C}_{\mathrm{m}} \in \mathbb{R}^{T \times s_{\mathrm{m}}}$. We denote a matrix with orthonormal columns by $\boldsymbol{U}_{\mathrm{m}}^{\perp} \in \mathbb{R}^{T \times (T - s_{\mathrm{m}})}$ which satisfies

$$\boldsymbol{U}_{\mathrm{m}} \boldsymbol{U}_{\mathrm{m}}^{\intercal} + \boldsymbol{U}_{\mathrm{m}}^{\perp} (\boldsymbol{U}_{\mathrm{m}}^{\perp})^{\intercal} = \boldsymbol{I}_T \quad \text{and} \quad \boldsymbol{U}_{\mathrm{m}}^{\intercal} \boldsymbol{U}_{\mathrm{m}}^{\perp} = \boldsymbol{O}.$$

Then,

$$\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}^{\perp})^{\intercal}\boldsymbol{K} = \boldsymbol{K} - \boldsymbol{U}_{\mathrm{m}}\boldsymbol{U}_{\mathrm{m}}^{\intercal}\boldsymbol{K} = \boldsymbol{K} - \boldsymbol{C}_{\mathrm{m}}\boldsymbol{C}_{\mathrm{m}}^{\dagger}\boldsymbol{K}.$$

The Count Sketch matrix has been used as a specific random projection technique due to the unbiasedness and efficiency while approximating the inner product of two vectors.

Lemma 1 (Lemma 2 from (Pham & Pagh, 2013))**.** *Given two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^T$, we denote by $\boldsymbol{S} \in \mathbb{R}^{T\times s}$ the Count Sketch matrix. Then,*

$$\begin{aligned}\mathrm{E}\left[\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle\right] &= \langle \boldsymbol{x}, \boldsymbol{y}\rangle,\\ \mathrm{Var}[\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle] &\leq \frac{1}{s}\left(\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2 + \|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2\right).\end{aligned}$$

We first provide the analysis of the expectation and variance while approximating the inner product using $\boldsymbol{S}_{\mathrm{p}}$ in the proposed incremental randomized sketches.

Lemma 2 (Inner Product Preserving Property)**.** *Given two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^T$, we denote by $\boldsymbol{S} \in \mathbb{R}^{T\times s}$ the sketch matrix $\boldsymbol{S}_{\mathrm{p}}$ in the proposed incremental randomized sketches. Then,*

$$\begin{aligned}\mathrm{E}\left[\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle\right] &= \langle \boldsymbol{x}, \boldsymbol{y}\rangle,\\ \mathrm{Var}[\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle] &\leq \frac{1}{s}\left(\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2 + \|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2\right).\end{aligned}$$

Proof. Since

$$\boldsymbol{S}^{\intercal}\boldsymbol{x} = [\boldsymbol{S}_1^{\intercal}\boldsymbol{x}; \cdots; \boldsymbol{S}_d^{\intercal}\boldsymbol{x}],$$

we have

$$\begin{aligned}&\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle\\ &= [(\boldsymbol{S}_1^{\intercal}\boldsymbol{x})^{\intercal}, \ldots, (\boldsymbol{S}_d^{\intercal}\boldsymbol{x})^{\intercal}][\boldsymbol{S}_1^{\intercal}\boldsymbol{y}; \cdots; \boldsymbol{S}_d^{\intercal}\boldsymbol{y}]\\ &= \sum_{i=1}^{d}\langle \boldsymbol{S}_i^{\intercal}\boldsymbol{x}, \boldsymbol{S}_i^{\intercal}\boldsymbol{y}\rangle.\end{aligned} \tag{3}$$

Let $\widetilde{\boldsymbol{S}}_i = \sqrt{d}\boldsymbol{S}_i$ that is a Count Sketch matrix. By (3) and Lemma 1, we have

$$\begin{aligned}\mathrm{E}[\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle] &= \sum_{i=1}^{d}\mathrm{E}[\langle \boldsymbol{S}_i{}^{\intercal}\boldsymbol{x}, \boldsymbol{S}_i{}^{\intercal}\boldsymbol{y}\rangle]\\ &= \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}[\langle \widetilde{\boldsymbol{S}}_i^{\intercal}\boldsymbol{x}, \widetilde{\boldsymbol{S}}_i^{\intercal}\boldsymbol{y}\rangle] = \langle \boldsymbol{x}, \boldsymbol{y}\rangle.\end{aligned}$$

Similarly, we have

$$\mathrm{Var}[\langle \boldsymbol{S}^{\intercal}\boldsymbol{x}, \boldsymbol{S}^{\intercal}\boldsymbol{y}\rangle] = \frac{1}{d^2}\sum_{i=1}^{d}\mathrm{Var}[\langle \widetilde{\boldsymbol{S}}_i^{\intercal}\boldsymbol{x}, \widetilde{\boldsymbol{S}}_i^{\intercal}\boldsymbol{y}\rangle]$$

and then the result for the variance holds by Lemma 1. □

Then we demonstrate the unbiasedness of $\boldsymbol{S}_{\mathrm{p}}$ while approximating matrix products, which shows that our sketched kernel matrix approximation problem is an unbiased estimate of the modified Nyström kernel matrix approximation problem.

Lemma 3 (Matrix Product Preserving Property)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{T\times m}$, $\boldsymbol{B} \in \mathbb{R}^{p\times T}$. If $\boldsymbol{S} \in \mathbb{R}^{T\times s}$ is the sketch matrix $\boldsymbol{S}_{\mathrm{p}}$ in the proposed incremental randomized sketches, then,*

1) $\mathrm{E}\left[\|\boldsymbol{S}^{\intercal}\boldsymbol{A}\|_{\mathrm{F}}^2\right] = \|\boldsymbol{A}\|_{\mathrm{F}}^2, \quad \mathrm{Var}\left[\|\boldsymbol{S}^{\intercal}\boldsymbol{A}\|_{\mathrm{F}}^2\right] \leq \frac{2}{s}\|\boldsymbol{A}\|_{\mathrm{F}}^4.$

2) $\mathrm{E}\left[\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^{\intercal}\boldsymbol{A}\right] = \boldsymbol{B}\boldsymbol{A}.$

3) $\mathrm{E}\left[\|\boldsymbol{B}\boldsymbol{A} - \boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^{\intercal}\boldsymbol{A}\|_{\mathrm{F}}^2\right] \leq \frac{2}{s}\|\boldsymbol{B}\|_{\mathrm{F}}^2\|\boldsymbol{A}\|_{\mathrm{F}}^2.$

Proof. 1) It is obvious that

$$\mathrm{E}\left[\|\boldsymbol{S}^{\intercal}\boldsymbol{A}\|_{\mathrm{F}}^2\right] = \mathrm{E}\left[\sum_{i=1}^{m}X_i\right] = \sum_{i=1}^{m}\mathrm{E}\left[X_i\right],$$

where $X_i = \|\boldsymbol{S}^{\intercal}[\boldsymbol{A}]_{*i}\|_2^2$.

Using Lemma 2, we have

$$\mathrm{E}\left[X_i\right] = \|[\boldsymbol{A}]_{*i}\|_2^2,$$

and then

$$\mathrm{E}\left[\|\boldsymbol{S}^{\intercal}\boldsymbol{A}\|_{\mathrm{F}}^2\right] = \sum_{i=1}^{m}\|[\boldsymbol{A}]_{*i}\|_2^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2.$$

We can observe that

$$\begin{aligned}\mathrm{E}\left[\|\boldsymbol{S}^{\intercal}\boldsymbol{A}\|_{\mathrm{F}}^4\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{m}X_i\right)^2\right]\\ &= \mathrm{E}\left[\sum_{i,j=1}^{m}X_iX_j\right]\\ &= \sum_{i,j=1}^{m}\mathrm{E}\left[X_iX_j\right].\end{aligned} \tag{4}$$

By Lemma 2, we obtain

$$\mathrm{Var}\left[X_i\right] \leq \frac{2}{s}\|[\boldsymbol{A}]_{*i}\|_2^4,$$

and thus

$$\begin{aligned}\mathrm{E}\left[X_iX_j\right] &= \mathrm{Cov}\left[X_i, X_j\right] + \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_j\right]\\ &\leq \sqrt{\mathrm{Var}\left[X_i\right]\mathrm{Var}\left[X_j\right]} + \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_j\right]\\ &= \frac{s+2}{s}\|[\boldsymbol{A}]_{*i}\|_2^2\|[\boldsymbol{A}]_{*j}\|_2^2.\end{aligned} \tag{5}$$

From (4) and (5), we have

$$\begin{aligned}\mathrm{E}\left[\|\boldsymbol{S}^\intercal\boldsymbol{A}\|_\mathrm{F}^4\right] &\le \frac{s+2}{s}\sum_{i,j=1}^{m}\left(\|[\boldsymbol{A}]_{*i}\|_2^2\,\|[\boldsymbol{A}]_{*j}\|_2^2\right)\\ &= \frac{s+2}{s}\|\boldsymbol{A}\|_\mathrm{F}^4.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathrm{Var}\left[\|\boldsymbol{S}^\intercal\boldsymbol{A}\|_\mathrm{F}^2\right] &= \mathrm{E}\left[\|\boldsymbol{S}^\intercal\boldsymbol{A}\|_\mathrm{F}^4\right] - \mathrm{E}^2\left[\|\boldsymbol{S}^\intercal\boldsymbol{A}\|_\mathrm{F}^2\right]\\ &\le \frac{s+2}{s}\|\boldsymbol{A}\|_\mathrm{F}^4 - \|\boldsymbol{A}\|_\mathrm{F}^4\\ &= \frac{2}{s}\|\boldsymbol{A}\|_\mathrm{F}^4.\end{aligned}$$

2) Since

$$[\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{A}]_{ij} = [\boldsymbol{B}]_{i*}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal[\boldsymbol{A}]_{*j} = \langle\boldsymbol{S}^\intercal[\boldsymbol{B}]_{i*}, \boldsymbol{S}^\intercal[\boldsymbol{A}]_{*j}\rangle,$$

by Lemma 2 we can obtain

$$\mathrm{E}\left[[\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{A}]_{ij}\right] = \langle[\boldsymbol{B}]_{i*}, [\boldsymbol{A}]_{*j}\rangle = [\boldsymbol{B}\boldsymbol{A}]_{ij}. \quad (6)$$

3) Let $\boldsymbol{Y}_{ij} = [\boldsymbol{B}\boldsymbol{A} - \boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{A}]_{ij}$. By (6) we have $\mathrm{E}\left[\boldsymbol{Y}_{ij}\right] = 0$. Then,

$$\mathrm{E}\left[\boldsymbol{Y}_{ij}^2\right] = \mathrm{Var}\left[\boldsymbol{Y}_{ij}\right] + \mathrm{E}^2\left[\boldsymbol{Y}_{ij}\right] = \mathrm{Var}\left[\boldsymbol{Y}_{ij}\right].$$

By Lemma 2 and the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{Y}_{ij}^2\right] &= \mathrm{Var}\left[\boldsymbol{Y}_{ij}\right]\\ &= \mathrm{Var}\left[[\boldsymbol{B}\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{A}]_{ij}\right]\\ &= \mathrm{Var}\left[\langle\boldsymbol{S}^\intercal[\boldsymbol{B}]_{i*}, \boldsymbol{S}^\intercal[\boldsymbol{A}]_{*j}\rangle\right]\\ &\le \frac{1}{s}\left(\langle[\boldsymbol{B}]_{i*}, [\boldsymbol{A}]_{*j}\rangle^2 + \|[\boldsymbol{B}]_{i*}\|_2^2\|[\boldsymbol{A}]_{*j}\|_2^2\right)\\ &\le \frac{2}{s}\|[\boldsymbol{B}]_{i*}\|_2^2\|[\boldsymbol{A}]_{*j}\|_2^2.\end{aligned}$$

Thus

$$\mathrm{E}\left[\sum_{ij}\boldsymbol{Y}_{ij}^2\right] = \sum_{ij}\mathrm{E}\left[\boldsymbol{Y}_{ij}^2\right] \le \frac{2}{s}\|\boldsymbol{B}\|_\mathrm{F}^2\|\boldsymbol{A}\|_\mathrm{F}^2.$$

□

The following lemma provides that all singular values of $\boldsymbol{S}^\intercal\boldsymbol{U}$ lie in $[1-\epsilon_0, 1+\epsilon_0]$ with high probability.

Lemma 4 ((Nelson & Nguyên, 2013)). *Let $\boldsymbol{U} \in \mathbb{R}^{T\times c}$ be a matrix with orthonormal columns, $\boldsymbol{S} \in \mathbb{R}^{T\times s}$ the sketch matrix $\boldsymbol{S}_\mathrm{p}$ in the proposed incremental randomized sketches. Set $d = \Theta(\epsilon_0^{-1}\log^3(c\delta_0^{-1}))$ for $\boldsymbol{S}$. For $\epsilon_0 \in (0,1)$, with probability at least $1-\delta_0$ all singular values of $\boldsymbol{S}^\intercal\boldsymbol{U}$*

$$\sigma_i(\boldsymbol{S}^\intercal\boldsymbol{U}) = 1 \pm \epsilon_0$$

as long as

$$s \ge \frac{c\log^8\left(c\delta_0^{-1}\right)}{\epsilon_0^2}.$$

Further, this holds if the hash function h and σ defining the $\boldsymbol{S}$ is $\Omega\left(\log(c\delta_0^{-1})\right)$-wise independent.

Then we demonstrate that the proposed incremental randomized sketching is nearly as accurate as the modified Nyström for kernel matrix approximation, which shows that the proposed incremental randomized sketching achieves a relative-error bound for kernel matrix approximation, for the modified Nyström is a $1+\epsilon'$ relative-error approximation with respect to the best rank-k approximation.

Theorem 1 (Low-Rank Approximation Property). *Let $\boldsymbol{K} \in \mathbb{R}^{T\times T}$ be a symmetric matrix, $\epsilon_0 \in (0,1)$. $\boldsymbol{F}_\mathrm{sk} \in \mathbb{R}^{s_\mathrm{m}\times s_\mathrm{m}}$, $\boldsymbol{C}_\mathrm{m} \in \mathbb{R}^{T\times s_\mathrm{m}}$ are matrices defined in (1). If $\boldsymbol{S}_\mathrm{p} \in \mathbb{R}^{T\times s_\mathrm{p}}$ is the sketch matrix in the proposed incremental randomized sketches with $d = \Theta\left(\log^3(s_\mathrm{m})\right)$, let $\tau = s_\mathrm{m}/s_\mathrm{p}$ and assume*

$$s_\mathrm{p} = \Omega\left(s_\mathrm{m}\,\mathrm{polylog}\left(s_\mathrm{m}\delta_0^{-1}\right)/\epsilon_0^2\right),$$

then with probability at least $1-\delta_0$ all singular values of $\boldsymbol{S}_\mathrm{p}^\intercal\boldsymbol{U}_\mathrm{m}$ are $1\pm\epsilon_0$, and with probability at least $1-\delta$

$$\|\boldsymbol{C}_\mathrm{m}\boldsymbol{F}_\mathrm{sk}\boldsymbol{C}_\mathrm{m}^\intercal - \boldsymbol{K}\|_\mathrm{F}^2 \le (1+\epsilon)\|\boldsymbol{C}_\mathrm{m}\boldsymbol{F}_\mathrm{mod}\boldsymbol{C}_\mathrm{m}^\intercal - \boldsymbol{K}\|_\mathrm{F}^2,$$

where

$$\sqrt{\epsilon} = 2\tau\sqrt{\frac{T}{\delta_1\delta_2}} + \sqrt{\frac{2\tau}{\delta_2}}\left(\epsilon_0^2 + 2\epsilon_0 + 2\right),$$

δ_i is the failure probability of matrix product preserving as

$$\Pr\left\{\frac{\|\boldsymbol{B}_i\boldsymbol{A}_i - \boldsymbol{B}_i\boldsymbol{S}_\mathrm{p}\boldsymbol{S}_\mathrm{p}{}^\intercal\boldsymbol{A}_i\|_\mathrm{F}^2}{\|\boldsymbol{B}_i\|_\mathrm{F}^2\|\boldsymbol{A}_i\|_\mathrm{F}^2} > \frac{2}{\delta_i s_\mathrm{p}}\right\} \le \delta_i,\ \ i=1,2,$$

$\boldsymbol{A}_1 = \boldsymbol{U}_\mathrm{m}$, $\boldsymbol{B}_1 = \boldsymbol{I}_T$, $\boldsymbol{A}_2 = \boldsymbol{U}_\mathrm{m}^\perp(\boldsymbol{U}_\mathrm{m}{}^\perp)^\intercal\boldsymbol{K}$, $\boldsymbol{B}_2 = \boldsymbol{U}_\mathrm{m}^\intercal$, and $\delta = \delta_0 + \delta_1 + \delta_2$.

Proof. Let $\boldsymbol{A} \in \mathbb{R}^{T\times m}$, $\boldsymbol{B} \in \mathbb{R}^{p\times T}$. By 3) in Lemma 3 and the Markov's inequality, with probability at least $1-\hat{\delta}$

$$\|\boldsymbol{B}\boldsymbol{A} - \boldsymbol{B}\boldsymbol{S}_\mathrm{p}\boldsymbol{S}_\mathrm{p}{}^\intercal\boldsymbol{A}\|_\mathrm{F}^2 \le \frac{2}{\hat{\delta}s_\mathrm{p}}\|\boldsymbol{B}\|_\mathrm{F}^2\|\boldsymbol{A}\|_\mathrm{F}^2. \quad (7)$$

From (7) with probability at least $1-\delta_1$

$$\begin{aligned}\|\boldsymbol{I}_T\boldsymbol{S}_\mathrm{p}\boldsymbol{S}_\mathrm{p}^\intercal\boldsymbol{U}_\mathrm{m}\|_2 &\le \|\boldsymbol{I}_T\boldsymbol{S}_\mathrm{p}\boldsymbol{S}_\mathrm{p}^\intercal\boldsymbol{U}_\mathrm{m} - \boldsymbol{I}_T\boldsymbol{U}_\mathrm{m}\|_2 + \|\boldsymbol{I}_T\boldsymbol{U}_\mathrm{m}\|_2\\ &\le \|\boldsymbol{I}_T\boldsymbol{S}_\mathrm{p}\boldsymbol{S}_\mathrm{p}^\intercal\boldsymbol{U}_\mathrm{m} - \boldsymbol{I}_T\boldsymbol{U}_\mathrm{m}\|_\mathrm{F} + 1\\ &\le \sqrt{\frac{2}{\delta_1 s_\mathrm{p}}}\|\boldsymbol{I}_T\|_\mathrm{F}\|\boldsymbol{U}_\mathrm{m}\|_\mathrm{F} + 1\\ &= \sqrt{\frac{2Ts_\mathrm{m}}{\delta_1 s_\mathrm{p}}} + 1,\end{aligned}$$

and with probability at least $1-\delta_2$

$$\begin{aligned}
&\|(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})^{\intercal}\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}}\\
&=\|(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})^{\intercal}\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}{}^{\perp})^{\intercal}\boldsymbol{K}-\boldsymbol{U}_{\mathrm{m}}^{\intercal}\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}{}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}}\\
&\leq\sqrt{\frac{2}{\delta_2 s_{\mathrm{p}}}}\|\boldsymbol{U}_{\mathrm{m}}\|_{\mathrm{F}}\|\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}{}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}}\\
&=\sqrt{\frac{2s_{\mathrm{m}}}{\delta_2 s_{\mathrm{p}}}}\|\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}{}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}}.
\end{aligned}$$

Assume $s_{\mathrm{p}}\geq s_{\mathrm{m}}\log^8(s_{\mathrm{m}}\delta_0^{-1})/\epsilon_0^2$. By Lemma 4, with probability at least $1-\delta_0$

$$\sigma_{\max}^2(\boldsymbol{S}_{\mathrm{p}}\boldsymbol{U}_{\mathrm{m}})\leq(1+\epsilon_0)^2.$$

Let $\boldsymbol{E}=\boldsymbol{C}_{\mathrm{m}}\boldsymbol{F}_{\mathrm{mod}}\boldsymbol{C}_{\mathrm{m}}^{\intercal}-\boldsymbol{K}$. According to the above discussion, by Lemma 17 from (Wang et al., 2015), with probability at least $1-(\delta_0+\delta_1+\delta_2)$

$$\begin{aligned}
&\|(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})^{\intercal}\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{E}\boldsymbol{S}_{\mathrm{p}}(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})\|_{\mathrm{F}}\\
&\leq\left(\|\boldsymbol{I}_T\boldsymbol{S}_{\mathrm{p}}\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}}\|_2+\sigma_{\max}^2(\boldsymbol{S}_{\mathrm{p}}\boldsymbol{U}_{\mathrm{m}})\right)\\
&\quad\|(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})^{\intercal}\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}}\\
&\leq\sqrt{\epsilon}\|\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}{}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}},
\end{aligned}$$

where

$$\sqrt{\epsilon}=2\tau\sqrt{\frac{T}{\delta_1\delta_2}}+\sqrt{\frac{2\tau}{\delta_2}}\left(\epsilon_0^2+2\epsilon_0+2\right).$$

It finally follows from Lemma 17 in (Wang et al., 2015) that with probability at least $1-(\delta_0+\delta_1+\delta_2)$

$$\begin{aligned}
&\|\boldsymbol{C}_{\mathrm{m}}\boldsymbol{F}_{\mathrm{sk}}\boldsymbol{C}_{\mathrm{m}}^{\intercal}-\boldsymbol{K}\|_{\mathrm{F}}^2\\
&\leq\|\boldsymbol{E}\|_{\mathrm{F}}^2+\sigma_{\min}^{-8}(\boldsymbol{S}_{\mathrm{p}}\boldsymbol{U}_{\mathrm{m}})\|(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})^{\intercal}\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{E}\boldsymbol{S}_{\mathrm{p}}(\boldsymbol{S}_{\mathrm{p}}^{\intercal}\boldsymbol{U}_{\mathrm{m}})\|_{\mathrm{F}}^2\\
&\leq\|\boldsymbol{E}\|_{\mathrm{F}}^2+\epsilon\|\boldsymbol{U}_{\mathrm{m}}^{\perp}(\boldsymbol{U}_{\mathrm{m}}{}^{\perp})^{\intercal}\boldsymbol{K}\|_{\mathrm{F}}^2\\
&=\|\boldsymbol{E}\|_{\mathrm{F}}^2+\epsilon\|\boldsymbol{K}-\boldsymbol{C}_{\mathrm{m}}\boldsymbol{C}_{\mathrm{m}}^{\dagger}\boldsymbol{K}\|_{\mathrm{F}}^2\\
&\leq(1+\epsilon)\|\boldsymbol{E}\|_{\mathrm{F}}^2.
\end{aligned}$$

□

Denote the hypothesis at round t using the updated explicit feature mapping $\boldsymbol{\phi}_{t+1}(\cdot)$ by

$$\bar{f}_t(\boldsymbol{x}_t)=\langle\bar{\boldsymbol{w}}_t,\boldsymbol{\phi}_{t+1}(\boldsymbol{x}_t)\rangle. \tag{8}$$

We can obtain $\bar{\boldsymbol{w}}_t$ by setting $\bar{f}_t(\boldsymbol{x}_t)=f_t(\boldsymbol{x}_t)$, which yields

$$\bar{\boldsymbol{w}}_t^{\intercal}=f_t(\boldsymbol{x}_t)\boldsymbol{\phi}_{t+1}(\boldsymbol{x}_t)^{\dagger}=f_t(\boldsymbol{x}_t)\frac{\boldsymbol{\phi}_{t+1}(\boldsymbol{x}_t)^{\intercal}}{\|\boldsymbol{\phi}_{t+1}(\boldsymbol{x}_t)\|_2^2}.$$

Then we update the hypothesis $\bar{f}_t(\cdot)=\langle\bar{\boldsymbol{w}}_t,\boldsymbol{\phi}_{t+1}(\cdot)\rangle$ in (8) as follows:

$$\boldsymbol{w}_{t+1}=\bar{\boldsymbol{w}}_t-\eta\nabla\mathcal{L}_t(\bar{\boldsymbol{w}}_t), \tag{9}$$

where

$$\ell_t(\bar{\boldsymbol{w}}_t)=\ell_t(\bar{f}_t)=\ell(\bar{f}_t(\boldsymbol{x}_t),y_t)$$

and

$$\mathcal{L}_t(\bar{\boldsymbol{w}}_t)=\ell_t(\bar{\boldsymbol{w}}_t)+\frac{\lambda}{2}\|\bar{\boldsymbol{w}}_t\|_2^2.$$

Let $\boldsymbol{K}_{B,\rho}\in\mathbb{R}^{(B+\lfloor(T-B)/\rho\rfloor)\times(B+\lfloor(T-B)/\rho\rfloor)}$ be the intersection matrix of $\boldsymbol{K}$ which is constructed by $B+\lfloor(T-B)/\rho\rfloor$ examples, $\mu(\boldsymbol{K}_{B,\rho})$ be the coherence of $\boldsymbol{K}_{B,\rho}$ as

$$\mu(\boldsymbol{K}_{B,\rho})=\frac{B+\lfloor(T-B)/\rho\rfloor}{\mathrm{rank}(\boldsymbol{K}_{B,\rho})}\max_i\|(\boldsymbol{U}_{B,\rho})_{i,:}\|_2^2,$$

where $\boldsymbol{U}_{B,\rho}$ is the singular vector matrix of $\boldsymbol{K}_{B,\rho}$. We finally obtain the following regret bound for online kernel learning[1].

Theorem 2 (Regret Bound). *Let $\boldsymbol{K}\in\mathbb{R}^{T\times T}$ be a kernel matrix with $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)\leq 1$, $\epsilon_0\in(0,1)$, δ_i $(i=0,1,2)$ be the failure probabilities defined in Theorem 1, and k $(k\leq s_{\mathrm{p}})$ be the rank in the incremental randomized sketches. Set the update cycle $\rho=\lfloor\theta(T-B)\rfloor$, $\theta\in(0,1)$,*

$$d=\Theta(\log^3(s_{\mathrm{m}}))\quad\textit{and}\quad\tau=s_{\mathrm{m}}/s_{\mathrm{p}},$$

for the sketch matrix $\boldsymbol{S}_{\mathrm{p}}\in\mathbb{R}^{T\times s_{\mathrm{p}}}$ in the proposed incremental randomized sketches. Assume ℓ_t is a convex loss function that is Lipschitz continuous with the Lipschitz constant L, and the eigenvalues of $\boldsymbol{K}$ decay polynomially with decay rate $\beta>1$. Let $\boldsymbol{w}_t$, $t\in[T]$ be the sequence of hypotheses generated by (9), satisfying

$$|f_t(\boldsymbol{x}_t)|=|\langle\boldsymbol{w}_t,\boldsymbol{\phi}_t(\boldsymbol{x}_t)\rangle|\leq C_f,\quad t\in[T].$$

For the optimal hypothesis f^* that minimizes

$$\mathcal{L}(f)=\frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2+\frac{1}{T}\sum_{t=1}^{T}\ell_t(f),$$

in the original reproducing kernel Hilbert space $\mathcal{H}_\kappa$, if

$$\begin{aligned}
s_{\mathrm{p}}&=\Omega\left(s_{\mathrm{m}}\,\mathrm{polylog}(s_{\mathrm{m}}\delta_0^{-1})/\epsilon_0^2\right),\\
s_{\mathrm{m}}&=\Omega(\mu(\boldsymbol{K}_{B,\rho})k\log k),
\end{aligned}$$

then with probability at least $1-\delta$

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\mathcal{L}_t(\boldsymbol{w}_t)-\mathcal{L}_t(f^*)\right)\\
&\leq\frac{C_f^2+\|\boldsymbol{w}_Z^*\|_2^2}{2\theta\eta}+\frac{\eta L^2T}{2}+\\
&\quad\frac{1}{\lambda(\beta-1)}\left(\frac{3}{2}-\frac{B+1/\theta}{T}\right)+\frac{\sqrt{1+\epsilon}}{\lambda}O(\sqrt{B}),
\end{aligned}$$

[1] In the theoretical analysis, we assume $T_0=B$ and omit KOGD used at the first stage that enjoys a $O(\sqrt{B})$ regret bound.

where $\delta = \delta_0 + \delta_1 + \delta_2$, $\boldsymbol{w}_t^*$ is the optimal hypothesis on the incremental randomized sketches over the first t instances and $Z = \arg\max_{i\in[T]} \|\boldsymbol{w}_i^*\|_2$,

$$\sqrt{\epsilon} = 2\tau\sqrt{\frac{T}{\delta_1\delta_2}} + \sqrt{\frac{2\tau}{\delta_2}}\left(\epsilon_0^2 + 2\epsilon_0 + 2\right).$$

Proof. By Theorem 2 in (Yang et al., 2012), we have

$$\mathcal{L}(\boldsymbol{w}^*) - \mathcal{L}(f^*) \leq \frac{1}{2T\lambda}\|\boldsymbol{K}_{\mathrm{sk}}^{(T)} - \boldsymbol{K}\|_2,$$

where $\boldsymbol{w}^*$ is the optimal hypothesis on the incremental randomized sketches in hindsight, which yields

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\mathcal{L}_t(\boldsymbol{w}^*) - \mathcal{L}_t(f^*)\right) \\
&\leq \frac{1}{2\lambda}\left\|\widehat{[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]}_{B,\rho} - \boldsymbol{K}\right\|_2 \\
&\leq \frac{1}{2\lambda}\left(\left\|\widehat{[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]}_{B,\rho} - \widehat{\boldsymbol{K}}_{B,\rho}\right\|_2 + \left\|\widehat{\boldsymbol{K}}_{B,\rho} - \boldsymbol{K}\right\|_2\right) \\
&= \frac{1}{2\lambda}\left(\left\|[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]_{B,\rho} - \boldsymbol{K}_{B,\rho}\right\|_2 + \left\|\widehat{\boldsymbol{K}}_{B,\rho} - \boldsymbol{K}\right\|_2\right)
\end{aligned} \tag{10}$$

where $\boldsymbol{K}_{B,\rho} \in \mathbb{R}^{(B+\lfloor(T-B)/\rho\rfloor)\times(B+\lfloor(T-B)/\rho\rfloor)}$ is the intersection matrix of $\boldsymbol{K}$ which is constructed by $B + \lfloor(T - B)/\rho\rfloor$ examples, $[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]_{B,\rho}$ is the approximate matrix for $\boldsymbol{K}_{B,\rho}$ using the proposed incremental randomized sketching with rank parameter k, $\boldsymbol{O}$ is a zero matrix of size $(T - B - \lfloor(T - B)/\rho\rfloor) \times (T - B - \lfloor(T - B)/\rho\rfloor)$ and

$$\begin{aligned}
\widehat{\boldsymbol{K}}_{B,\rho} &= \mathrm{diag}\left\{\boldsymbol{K}_{B,\rho},\, \boldsymbol{O}\right\} \in \mathbb{R}^{T\times T}, \\
\widehat{[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]}_{B,\rho} &= \mathrm{diag}\left\{[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]_{B,\rho},\, \boldsymbol{O}\right\} \in \mathbb{R}^{T\times T}.
\end{aligned}$$

Since the eigenvalues of the kernel matrix decay polynomially with decay rate $\beta > 1$, the following bound holds

$$\begin{aligned}
&\|\widehat{\boldsymbol{K}}_{B,\rho} - \boldsymbol{K}\|_2 \\
&\leq \frac{T - B - \lfloor(T - B)/\rho\rfloor}{T}\sum_{i=1}^{T} i^{-\beta} \\
&\leq \frac{T - B - \lfloor(T - B)/\rho\rfloor}{T}\int_1^T i^{-\beta}\mathrm{d}i \\
&= \frac{T - B - \lfloor(T - B)/\rho\rfloor}{T}\frac{1}{\beta - 1}\left(1 - \frac{1}{T^{\beta-1}}\right) \\
&\leq \frac{1}{\beta - 1}\left(1 - \frac{B + \lfloor(T - B)/\rho\rfloor}{T}\right).
\end{aligned} \tag{11}$$

From Theorem 1, with probability at least $1 - \delta$,

$$\begin{aligned}
&\left\|[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]_{B,\rho} - \boldsymbol{K}_{B,\rho}\right\|_2 \\
&\leq \sqrt{1+\epsilon}\,\|[\boldsymbol{C}_{\mathrm{m}}\boldsymbol{F}_{\mathrm{mod}}\boldsymbol{C}_{\mathrm{m}}^{\intercal}]_{B,\rho} - \boldsymbol{K}_{B,\rho}\|_{\mathrm{F}},
\end{aligned} \tag{12}$$

where $[\boldsymbol{C}_{\mathrm{m}}\boldsymbol{F}_{\mathrm{mod}}\boldsymbol{C}_{\mathrm{m}}^{\intercal}]_{B,\rho}$ is the approximate matrix for $\boldsymbol{K}_{B,\rho}$ using the modified Nyström approach with rank parameter k.

Denote the best rank-k approximation of $\boldsymbol{A}$ by $(\boldsymbol{A})_k$. Since the eigenvalues of $\boldsymbol{K}$ decay polynomially with decay rate $\beta > 1$, there exists $\beta > 1$ such that $\lambda_i(\boldsymbol{K}) = O(i^{-\beta})$, which yields

$$\begin{aligned}
&\|\boldsymbol{K}_{B,\rho} - (\boldsymbol{K}_{B,\rho})_k\|_{\mathrm{F}} \\
&= \sqrt{B + \lfloor(T - B)/\rho\rfloor - k}\cdot(k+1)^{-\beta} \\
&= O(\sqrt{B}).
\end{aligned} \tag{13}$$

Given $\epsilon' \in (0, 1)$, when $s_{\mathrm{m}} = \Omega(\mu(\boldsymbol{K}_{B,\rho})k\log k)$, the following bound holds (Wang et al., 2016)

$$\begin{aligned}
&\|[\boldsymbol{C}_{\mathrm{m}}\boldsymbol{F}_{\mathrm{mod}}\boldsymbol{C}_{\mathrm{m}}^{\intercal}]_{B,\rho} - \boldsymbol{K}_{B,\rho}\|_{\mathrm{F}} \\
&\leq \sqrt{1+\epsilon'}\,\|\boldsymbol{K}_{B,\rho} - (\boldsymbol{K}_{B,\rho})_k\|_{\mathrm{F}},
\end{aligned} \tag{14}$$

where $\mu(\boldsymbol{K}_{B,\rho})$ is the coherence of $\boldsymbol{K}_{B,\rho}$. Combining (12), (13) with (14), we obtain

$$\left\|[\boldsymbol{K}_{\mathrm{sk}}^{(T)}]_{B,\rho} - \boldsymbol{K}_{B,\rho}\right\|_2 \leq \sqrt{1+\epsilon}\,O(\sqrt{B}). \tag{15}$$

Substituting (11) and (15) into (10), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\mathcal{L}_t(\boldsymbol{w}^*) - \mathcal{L}_t(f^*)\right) \\
&\leq \frac{1}{2\lambda(\beta-1)}\left(1 - \frac{B + \lfloor(T - B)/\rho\rfloor}{T}\right) + \\
&\frac{\sqrt{1+\epsilon}}{2\lambda}O(\sqrt{B}).
\end{aligned} \tag{16}$$

Then we analyze the regret caused by hypothesis updating on the incremental randomized sketches. We first decompose $\mathcal{L}_t(\boldsymbol{w}_t) - \mathcal{L}_t(\boldsymbol{w}^*)$ into two terms as follows:

$$\begin{aligned}
&\mathcal{L}_t(\boldsymbol{w}_t) - \mathcal{L}_t(\boldsymbol{w}^*) \\
&= \underbrace{\mathcal{L}_t(\boldsymbol{w}_t) - \mathcal{L}_t(\boldsymbol{w}_t^*)}_{\text{Optimization error}} + \underbrace{\mathcal{L}_t(\boldsymbol{w}_t^*) - \mathcal{L}_t(\boldsymbol{w}^*)}_{\text{Estimation error}},
\end{aligned}$$

where $f_t^*(\cdot) = \langle\boldsymbol{w}_t^*, \boldsymbol{\phi}_t(\cdot)\rangle$ is the optimal hypothesis on the incremental randomized sketches over the first t instances, and $\boldsymbol{w}^*$ is the optimal hypothesis on the incremental randomized sketches in hindsight. The optimization error measures the discrepancy between the hypothesis generated by our sketched online gradient descent and the optimal hypothesis on the incremental randomized sketches at each round, and the estimation error measures the difference between the optimal hypotheses on the incremental randomized sketches over the first t instances and all the T instances respectively. For the optimization error, by the convexity of loss function,

we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\mathcal{L}_t(\boldsymbol{w}_t)-\mathcal{L}_t(\boldsymbol{w}_t^*)\right)\\
\leq&\sum_{i=1}^{\lfloor(T-B)/\rho\rfloor}\frac{\|\bar{\boldsymbol{w}}_{B+(i-1)\rho}-\boldsymbol{w}_{B+i\rho}^*\|_2^2}{2\eta}+\frac{\eta L^2T}{2}\\
\leq&\sum_{i=1}^{\lfloor(T-B)/\rho\rfloor}\frac{\|\bar{\boldsymbol{w}}_{B+(i-1)\rho}\|_2^2+\|\boldsymbol{w}_{B+i\rho}^*\|_2^2}{2\eta}+\frac{\eta L^2T}{2}\\
\leq&\sum_{i=1}^{\lfloor(T-B)/\rho\rfloor}\frac{|f_{B+(i-1)\rho}(\boldsymbol{x}_{B+(i-1)\rho})|^2+\|\boldsymbol{w}_{B+i\rho}^*\|_2^2}{2\eta}+\\
&\frac{\eta L^2T}{2}\\
\leq&\left\lfloor\frac{T-B}{\rho}\right\rfloor\frac{C_f^2+\|\boldsymbol{w}_Z^*\|_2^2}{2\eta}+\frac{\eta L^2T}{2},
\end{aligned}\tag{17}$$

where $\bar{\boldsymbol{w}}_B=\boldsymbol{w}_{T_0+1}$, $Z=\arg\max_{i\in[T]}\|\boldsymbol{w}_i^*\|_2$ and $|f_t(\boldsymbol{x}_t)|\leq C_f, t\in[T]$. For the estimation error, we obtain the following upper bound

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\mathcal{L}_t(\boldsymbol{w}_t^*)-\mathcal{L}_t(\boldsymbol{w}^*)\right)\\
\leq&\frac{1}{2\lambda}\left\|\boldsymbol{K}_{\mathrm{sk}}^{(T_0)}-\boldsymbol{K}_{\mathrm{sk}}^{(T)}\right\|_2\\
\leq&\frac{1}{2\lambda}\left(\left\|\boldsymbol{K}_{\mathrm{sk}}^{(T_0)}-\boldsymbol{K}^{(T_0)}\right\|_2+\left\|\boldsymbol{K}^{(T_0)}-\boldsymbol{K}\right\|_2+\right.\\
&\left.\left\|\boldsymbol{K}_{\mathrm{sk}}^{(T)}-\boldsymbol{K}\right\|_2\right)\\
\leq&\frac{1}{2\lambda}\left[\sqrt{1+\tilde{\epsilon}}\,O(\sqrt{B})+\frac{1}{\beta-1}\left(1-\frac{B}{T}\right)+\right.\\
&\left.\left\|\boldsymbol{K}_{\mathrm{sk}}^{(T)}-\boldsymbol{K}\right\|_2\right].
\end{aligned}\tag{18}$$

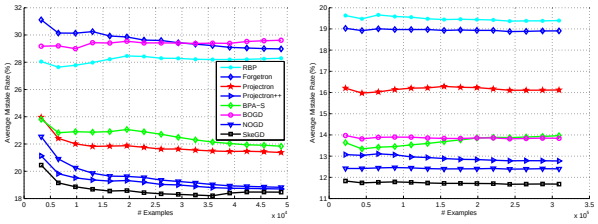
Finally, the three inequalities (16), (17) and (18) combined give the following bound

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\mathcal{L}_t(\boldsymbol{w}_t)-\mathcal{L}_t(f^*)\right)\\
\leq&\left\lfloor\frac{T-B}{\rho}\right\rfloor\frac{C_f^2+\|\boldsymbol{w}_Z^*\|_2^2}{2\eta}+\frac{\eta L^2T}{2}+\\
&\frac{1}{\lambda(\beta-1)}\left(\frac{3}{2}-\frac{B+\lfloor(T-B)/\rho\rfloor}{T}\right)+\\
&\frac{\sqrt{1+\epsilon}}{\lambda}O(\sqrt{B}).
\end{aligned}$$

□

3. More Experimental Results

To further analyze the convergence of the compared algorithms and SkeGD, we give the convergence curves in terms of the mistake rates on `a9a` and `cod-rna`. As the results shown in Figure 2, the mistake rates of our SkeGD converge much faster than the other online kernel learning algorithms, which demonstrates the efficiency and effectiveness of SkeGD. In Figure 2 (b), the growing mistake rates of the compared algorithms indicate that the buffer of support vectors cannot retain the key information with a small budget. Whereas SkeGD has a decreasing mistake rate, which demonstrates the effectiveness of the proposed incremental randomized sketches.



(a) `a9a` (b) `cod-rna`

Figure 2. The average mistake rates of RBP, Forgetron, Projectron, Projectron++, BPA-S, BOGD, NOGD and our SkeGD.

Table 1 shows the experimental results on the adversarial datasets `spambase-1` and `spambase-2`. As the same adversarial settings in `german-1` and `german-1`, we construct these two datasets using the benchmark `spambase`, and set $k_{\mathrm{b}}=500, k_{\mathrm{r}}=10$ in `spambase-1` and $k_{\mathrm{b}}=500$, $k_{\mathrm{r}}=20$ in `spambase-2`. From the results, we can observe that our SkeGD achieves the highest accuracy in adversarial environments, and is much more efficient than the second-order algorithm PROS-N-KONS, while having a comparable efficiency to the other first-order online kernel learning algorithms. Besides, an appropriate update cycle results in better performances with respect to the accuracy, which conforms to Remark 2.

Table 1. Comparison of online kernel learning algorithms in adversarial environments w.r.t. the mistake rates (%) and the running time (s), where $\rho=\lfloor\theta(T-B)\rfloor$ is the update cycle of SkeGD.

Algorithm	`spambase-1`		`spambase-2`	
	Mistake rate	Time	Mistake rate	Time
FOGD	39.793 ± 0.140	0.235	29.257 ± 0.186	0.478
NOGD	44.714 ± 0.001	0.468	40.586 ± 0.002	0.938
PROS-N-KONS	35.953 ± 0.423	265.26	26.800 ± 0.802	696.39
SkeGD ($\theta=0.1$)	29.034 ± 1.584	0.344	16.879 ± 3.779	0.646
SkeGD ($\theta=0.01$)	25.367 ± 1.134	0.494	$\mathbf{14.700\pm0.338}$	0.743
SkeGD ($\theta=0.005$)	$\mathbf{25.170\pm0.847}$	0.682	14.733 ± 2.562	0.848
SkeGD ($\theta=0.001$)	25.626 ± 0.966	2.337	14.845 ± 1.766	2.890

References

Nelson, J. and Nguyên, H. L. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pp. 117–126, 2013.

Pham, N. and Pagh, R. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 239–247, 2013.

Wang, S., Zhang, Z., and Zhang, T. Towards more efficient symmetric matrix sketching and the CUR matrix decomposition. *arXiv:1503.08395v4*, 2015.

Wang, S., Luo, L., and Zhang, Z. SPSD matrix approximation vis column selection: Theories, algorithms, and extensions. *Journal of Machine Learning Research*, 17: 1–49, 2016.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25*, pp. 476–484, 2012.