Signal Processing 199 (2022) 108603

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Passive sampling in reproducing kernel Hilbert spaces using leverage scores☆

Pere Giménez-Febrer[a,*], Alba Pagès-Zamora[a], Ignacio Santamaría[b]

[a] SPCOM Group, Universitat Politècnica de Catalunya-Barcelona Tech, Spain
[b] Dept. of Communications Engineering, Universidad de Cantabria, Spain

ARTICLE INFO

Article history:
Received 23 September 2021
Revised 13 March 2022
Accepted 23 April 2022
Available online 4 May 2022

Keywords:
Kernel ridge regression
Leverage score
Nyström approximation
Passive sampling
Reproducing kernel Hilbert space

ABSTRACT

This paper deals with the selection of the training dataset in kernel-based methods for function reconstruction, with a focus on kernel ridge regression. A functional analysis is performed which, in the absence of noise, links the optimal sampling distribution to the one minimizing the difference between the kernel matrix and its low-rank Nyström approximation. From this standpoint, a statistical passive sampling approach is derived which uses the leverage scores of the columns of the kernel matrix to design a sampling distribution that minimizes an upper bound of the risk function. The proposed approach constitutes a passive method, able to select the optimal subset of training samples using only information provided by the input set and the kernel, but without needing to know the values of the function to be approximated. Furthermore, the proposed approach is backed up by numerical tests on real datasets.

1. Introduction

Machine learning predicts labels by applying decision rules previously learned on a training dataset, which consists of a set of input samples and their labels. The training objective is to obtain a model that can generalize to previously unseen or unlabeled inputs. While in learning applications the focus is usually on achieving a small prediction error, which is heavily influenced by the choice of the training set, there is also the cost of label acquisition for training. This cost might be comparable or even superior to the prediction cost. For instance, in recommender systems, a request is sent to a user to rate an item; but this does not even guarantee a rating since the request might be ignored. In other areas such as sensor networks, medicine, or any scenario with human annotators, the time to label an input is also non-negligible. Hence, the design of optimal sample selection strategies is of paramount importance to improve algorithm performance both in terms of accuracy and cost. The existing sample selection strategies may be divided into active and passive, with the difference between them requiring labeled inputs or not.

Active sampling schemes allow to choose the most promising set of inputs to label. They implement an iterative scheme that, given the availability of a small starting training set, executes two steps repeatedly: first, the function is learned or updated using the current training set and, second, a criterion is evaluated to decide which should be the next input to be labeled and then added to the training set. There exist a variety of criteria for active sampling, the most common being based on prediction uncertainty. For instance, in Gaussian processes [1,2], the predictive variance serves as an uncertainty measure.

While active sampling schemes generally provide good results, their performance may be degraded when the samples are too noisy or when online operation is not possible [3]. There are also batch versions of active sampling which acquire the labels in groups, but the issue with noisy samples still remains. Moreover, most active learning methods are designed for classification tasks [4,5], which often have built-in uncertainty measures, while the availability of methods for regression is more limited [6–8].

An alternative to active sampling is provided by the passive sampling schemes. In passive sampling, the set of inputs to be labeled is selected by observing the geometry of the input space only. For instance, the greedy sampling approach in [3] iteratively adds new input samples to the training set by choosing the one with the largest distance to the set. By relying only on the input space, passive sampling avoids iteratively recomputing the learned function for every additional input. The selected set of inputs can be labeled all at once when it is deemed complete. Furthermore, the impact of noisy samples is diminished.

☆ This work has been funded by the Ministerio de Ciencia e Innovación (MICINN) of the Spanish Government and by the Agencia Estatal de Investigación (AEI/10.13039/501100011033) and ERDF funds (PID 2019-104958RB-C41/C43, RED2018-102668-T); and by the Catalan Government (2017 SGR 578).

* Corresponding author.
E-mail address: p.gimenez@upc.edu (P. Giménez-Febrer).

https://doi.org/10.1016/j.sigpro.2022.108603
0165-1684/

This paper addresses passive sampling for function approximation in a reproducing kernel Hilbert space (RKHS), with a focus on kernel ridge regression (KRR) [9–11]. To do so, a functional analysis is conducted which connects the concept of optimal sampling to the Nyström approximation [12] - a low-rank approximation to the kernel matrix built from a subset of its columns. Building atop this finding, a sample selection approach is presented based on selecting the samples by means of the leverage scores [13] of the kernel matrix columns. The proposed method enables passive batch sampling, hence selecting the entire training set without needing any label; only the kernel is needed. Moreover, the use of kernels enables the direct application of passive sampling to any input space, e.g., non-metric spaces, provided that a kernel function exists for it.

The paper is organized as follows. Section 2 provides a theoretical analysis of the sampling procedure in RKHSs and proves that Nyström-based passive sampling is optimal, in the absolute error sense, for noiseless samples. Section 3 extends the analysis to KRR with noisy samples. Section 4 introduces the passive sampling approach for KRR based on the leverage scores of the kernel matrix. Finally, Section 5 presents numerical tests and Section 6 offers conclusions.

Notation. Boldface lower-case fonts denote column vectors, boldface upper-case fonts denote matrices, and calligraphic upper-case letters denote sets. The ith entry of a vector $\boldsymbol{f}$ is denoted by $\boldsymbol{f}(i)$, the ith column of a matrix $\boldsymbol{K}$ by $\boldsymbol{K}(i)$, and the entry at row i column j by $\boldsymbol{K}_{ij}$. The superscript ^{T} denotes the transpose, † the pseudo inverse, and the hat $\hat{}$ is used for estimates. The symbol $\boldsymbol{I}$ stands for the identity matrix of appropriate size, specified by the context, and the trace operator is $\mathrm{Tr}(\cdot)$.

2. Optimal sampling and reconstruction in RKHSs

Let $f:\mathcal{X}\to\mathbb{R}$ be a function defined on an RKHS $\mathcal{H}_N$, and $\bar{\mathcal{X}}=\{\bar{x}_i\}_{i=1}^s$ be a set of sampled points in $\mathcal{X}$ where $|\mathcal{X}|=N$. Here, we focus on the problem of reconstructing f given a set of noiseless observations $\{\bar{y}_i\}_{i=1}^s$ such that $\bar{y}_i=f(\bar{x}_i)\ \forall i=1,\dots,s$.

The space $\mathcal{H}_N$ is a Hilbert space of $\mathbb{R}$-valued functions on a non-empty set $\mathcal{X}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_N}:\mathcal{H}_N\times\mathcal{H}_N\to\mathbb{R}$ and induced norm $\|f\|_{\mathcal{H}_N}=\langle f,f\rangle_{\mathcal{H}_N}$. A function $\kappa:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a reproducing kernel of $\mathcal{H}_N$, and $\mathcal{H}_N$ is a RKHS if

- $\forall x\in\mathcal{X},\kappa(\cdot,x)\in\mathcal{H}_N$,
- $\forall x\in\mathcal{X},\forall f\in\mathcal{H}_N,\langle f(\cdot),\kappa(\cdot,x)\rangle_{\mathcal{H}_N}=f(x)$,

where the latter is the reproducing property. In particular, the reproducing property implies that $\langle\kappa(\cdot,x),\kappa(\cdot,x')\rangle_{\mathcal{H}_N}=\kappa(x,x')$. In view of the Moore-Aronszajn theorem, one can construct the RKHS $\mathcal{H}_N$ as the completion of the space of functions spanned by the set $\{\kappa(\cdot,x_i)\}_{i=1}^N$, i.e.,

$$\mathcal{H}_N:=\{f:f(x)=\sum_{i=1}^N\alpha_i\kappa(x,x_i),\quad \alpha_i\in\mathbb{R}\} \tag{1}$$

with an inner product given by $\langle f,f'\rangle_{\mathcal{H}_N}=\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha'_j\kappa(x_i,x_j)$, where $\{\alpha_i\}_{i=1}^N$ and $\{\alpha'_j\}_{i=1}^N$ are the coefficients of $f,f'\in\mathcal{H}_N$, respectively. We refer to κ as the kernel function that spans $\mathcal{H}_N$, and to $\boldsymbol{K}$ as its kernel matrix with $\boldsymbol{K}_{ij}=\kappa(x_i,x_j)$. Note that our RKHS definition (1) differs from the usual one with an infinite-dimensional RKHS since $|\mathcal{X}|=N$ and, therefore, $\mathcal{H}_N$ is a finite-dimensional space [9].

In order to motivate the passive sampling approach introduced in Section 4, this section builds on the functional analysis approach to the signal reconstruction process from [14] and derives results valid for any algorithm operating in a RKHS.

The recovery of f from the observation set can be thought of as the "inversion" of

$$\bar{\boldsymbol{f}}=\mathcal{A}f \tag{2}$$

where $\bar{\boldsymbol{f}}=[f(\bar{x}_1),\dots,f(\bar{x}_s)]^T$, and $\mathcal{A}:\mathcal{H}_N\to\mathbb{R}^s$ is a sampling operator evaluating f at $\{\bar{x}_i\}_{i=1}^s$. Leveraging the reproducing property, the sampling operator is defined as

$$\mathcal{A}f=\sum_{n=1}^s\langle f,\kappa(\cdot,\bar{x}_n)\rangle_{\mathcal{H}_N}\boldsymbol{e}_n \tag{3}$$

where $\boldsymbol{e}_n$ is the nth vector of the standard basis, i.e., the nth column of the $s\times s$ identity matrix. Let $\mathcal{H}_s\subseteq\mathcal{H}_N$ be the subspace of dimension s spanned by $\{\kappa(\cdot,\bar{x}_n)\}_{n=1}^s$. Then, $f=f_s+f_s^\perp$, where f_s is the projection of f onto $\mathcal{H}_s$ and $f_s^\perp$ is the projection onto the subspace orthogonal to $\mathcal{H}_s$, denoted by $\mathcal{H}_s^\perp\subset\mathcal{H}_N$, such that $\langle f_s^\perp,\kappa(\cdot,\bar{x})\rangle_{\mathcal{H}_N}=0\ \forall\bar{x}\in\bar{\mathcal{X}}$. Thus, for $\bar{x}\in\bar{\mathcal{X}}$,

$$\begin{aligned}f(\bar{x})&=f_s(\bar{x})+f_s^\perp(\bar{x})=\langle f_s,\kappa(\cdot,\bar{x})\rangle_{\mathcal{H}_N}+\langle f_s^\perp,\kappa(\cdot,\bar{x})\rangle_{\mathcal{H}_N}\\&=\langle f_s,\kappa(\cdot,\bar{x})\rangle_{\mathcal{H}_N}=f_s(\bar{x}).\end{aligned} \tag{4}$$

Since $f_s\in\mathcal{H}_s$, it can be written as $\sum_{n=1}^s\bar{\alpha}_n\kappa(\cdot,\bar{x}_n)$ for some real coefficients $\{\bar{\alpha}_n\}_{n=1}^s$. In turn, $f(\bar{x})=\sum_{n=1}^s\bar{\alpha}_n\kappa(\bar{x},\bar{x}_n)$. Therefore, since $\bar{\boldsymbol{f}}(i):=f(\bar{x}_i)$, $\mathcal{A}$ maps f into a new RKHS $\overline{\mathcal{H}}_s\subseteq\mathbb{R}^s$ defined as

$$\overline{\mathcal{H}}_s:=\{\bar{\boldsymbol{f}}:\bar{\boldsymbol{f}}(i)=\sum_{n=1}^s\bar{\alpha}_n\kappa(\bar{x}_i,\bar{x}_n),\ \ \bar{\alpha}_n\in\mathbb{R}\} \tag{5}$$

with kernel matrix $\bar{\boldsymbol{K}}\in\mathbb{R}^{s\times s}$ where $\bar{\boldsymbol{K}}_{i,j}=\kappa(\bar{x}_i,\bar{x}_j)$ so that $\bar{\boldsymbol{f}}=\bar{\boldsymbol{K}}\bar{\boldsymbol{\alpha}}$, where $\bar{\boldsymbol{\alpha}}=[\bar{\alpha}_1,\dots,\bar{\alpha}_s]^T$, and $\langle\bar{\boldsymbol{f}},\bar{\boldsymbol{f}}'\rangle_{\overline{\mathcal{H}}_s}=\bar{\boldsymbol{f}}^T\bar{\boldsymbol{K}}^{-1}\bar{\boldsymbol{f}}'$ for $\bar{\boldsymbol{f}},\bar{\boldsymbol{f}}'\in\overline{\mathcal{H}}_s$ assuming that $\bar{\boldsymbol{K}}$ is invertible. With $\mathcal{A}:\mathcal{H}_N\to\overline{\mathcal{H}}_s$, its adjoint operator is $\mathcal{A}^*:\overline{\mathcal{H}}_s\to\mathcal{H}_N$, and by definition it satisfies for any $f''\in\mathcal{H}_N$ and $\bar{\boldsymbol{f}}'\in\overline{\mathcal{H}}_s$ that

$$\langle\mathcal{A}f'',\bar{\boldsymbol{f}}'\rangle_{\overline{\mathcal{H}}_s}=\langle f'',\mathcal{A}^*\bar{\boldsymbol{f}}'\rangle_{\mathcal{H}_N}. \tag{6}$$

Particularizing (6) for $f''=f$ and $\bar{\boldsymbol{f}}'=\bar{\boldsymbol{f}}$, with $\bar{\boldsymbol{f}}$ in (2), we obtain

$$\begin{aligned}\langle\mathcal{A}f,\bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}&=\langle\sum_{n=1}^s\langle f,\kappa(\cdot,\bar{x}_n)\rangle_{\mathcal{H}_N}\boldsymbol{e}_n,\bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}\\&=\sum_{n=1}^s\langle f,\kappa(\cdot,\bar{x}_n)\rangle_{\mathcal{H}_N}\langle\boldsymbol{e}_n,\bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}\\&=\langle f,\sum_{n=1}^s\langle\boldsymbol{e}_n,\bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}\kappa(\cdot,\bar{x}_n)\rangle_{\mathcal{H}_N}.\end{aligned} \tag{7}$$

This leads to the definition

$$\mathcal{A}^*\bar{\boldsymbol{f}}=\sum_{n=1}^s\langle\boldsymbol{e}_n,\bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}\kappa(\cdot,\bar{x}_n)=\sum_{n=1}^s\bar{\alpha}_n\kappa(\cdot,\bar{x}_n) \tag{8}$$

where $\langle\boldsymbol{e}_n,\bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}=\bar{\boldsymbol{f}}^T\bar{\boldsymbol{K}}^{-1}\boldsymbol{e}_n=\bar{\alpha}_n$. The operator $\mathcal{A}^*$ maps $\bar{\boldsymbol{f}}$ back into $\mathcal{H}_N$ using only the basis functions in $\mathcal{H}_s$, hence reversing the sampling in (2) and yielding an estimate for f. Moreover, the sampled values $f(\bar{x}_i)_{i=1}^s$ are not modified by $\mathcal{A}^*$. Hence, since $\mathcal{A}\mathcal{A}^*\bar{\boldsymbol{f}}=\bar{\boldsymbol{f}}$, we have that $\mathcal{A}\mathcal{A}^*=\mathcal{I}$; this can be seen by substituting (8) into (3).

After the sampling operator and its adjoint have been defined, the reconstruction of a function in $\mathcal{H}_N$ from its samples can be achieved through the consecutive application of each operator, i.e., $\hat{f}=\mathcal{A}^*\mathcal{A}f$. Hence, $\mathcal{P}=\mathcal{A}^*\mathcal{A}$ is an orthogonal projector onto $\mathcal{H}_s$ since it is a self-adjoint operator, i.e., $\langle\mathcal{P}f,f'\rangle_{\mathcal{H}_N}=\langle f,\mathcal{P}f'\rangle_{\mathcal{H}_N}$, and it satisfies $\mathcal{P}\mathcal{P}=\mathcal{P}$. Then, using (8) the estimate $\hat{f}$ is

$$\hat{f}=\mathcal{P}f=\mathcal{A}^*\bar{\boldsymbol{f}}=\sum_{n=1}^s\bar{\alpha}_n\kappa(\cdot,\bar{x}_n)=f_s. \tag{9}$$

Since an orthogonal projection onto a space yields an element with minimum distance to the original function, the reconstruction error at $x \in \mathcal{X}$ is upper bounded by

$$\begin{aligned}|f(x) - \mathcal{P}f(x)| &= |\langle f, \kappa(\cdot,x)\rangle_{\mathcal{H}_N} - \langle \mathcal{P}f, \kappa(\cdot,x)\rangle_{\mathcal{H}_N}| \\ &\overset{(a)}{=} |\langle f, \kappa(\cdot,x)\rangle_{\mathcal{H}_N} - \langle f, \mathcal{P}\kappa(\cdot,x)\rangle_{\mathcal{H}_N}| \\ &= |\langle f, \kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x)\rangle_{\mathcal{H}_N}| \\ &\overset{(b)}{\leq} \|f\|_{\mathcal{H}_N}\|\kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x)\|_{\mathcal{H}_N} \end{aligned} \tag{10}$$

where (a) uses the self-adjoint property of $\mathcal{P}$, and (b) uses the Cauchy-Schwarz inequality. It can be seen that the error upper bound at x depends on the distance between the kernel function $\kappa(\cdot,x)$ and its projection onto $\mathcal{H}_s$. Note that zero error is obtained at the sampled points since $\mathcal{P}\kappa(\cdot,\bar{x}) = \kappa(\cdot,\bar{x})\ \forall \bar{x} \in \bar{\mathcal{X}}$. For $x \notin \bar{\mathcal{X}}$, the error can only be zero if $\kappa(\cdot,x) \in \mathcal{H}_s$; this is the case when a kernel function is duplicated such that $\kappa(\cdot,\bar{x}_i) = \kappa(\cdot,x_j)$ where $\bar{x}_i \in \bar{\mathcal{X}}$ and $x_j \notin \bar{\mathcal{X}}$. This might happen, for instance, in a recommender system where two different users have identical tastes; the kernel function evaluated at each of the two users has identical value as well. Assuming no such duplicities exist,

$$|f(x) - \mathcal{P}f(x)| \leq \begin{cases} 0 & \text{if } x \in \bar{\mathcal{X}} \\ \|f\|_{\mathcal{H}_N}\|\kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x)\|_{\mathcal{H}_N} & \text{if } x \notin \bar{\mathcal{X}} \end{cases} \tag{11}$$

The absolute error over the function estimate is denoted by $|f - \hat{f}| = \sum_{i=1}^{N} |f(x_i) - \hat{f}(x_i)|$, and (11) shows that it can only be zero when $\hat{f} \in \mathcal{H}_s$ so that $\mathcal{P}f = f$. Therefore, in order to have zero error for any $f \in \mathcal{H}_N$, then $\mathcal{H}_N = \mathcal{H}_s$ must be true.

The second norm in the last inequality of (10) provides the distance in $\mathcal{H}_N$ between a kernel function and its closest approximation built with the functions in $\mathcal{H}_s$. Alternatively, this difference can be written in terms of $\mathcal{A}$ as shown in the lemma below, which will be used later.

Lemma 1. *The distance between $\kappa(\cdot,x)$ and its projection onto $\mathcal{H}_s\ \forall x \in \mathcal{X}$ is equal to the difference between the norm of $\kappa(\cdot,x)$ in $\mathcal{H}_N$ and its sampled counterpart in $\overline{\mathcal{H}}_s$,*

$$\|\kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x)\|_{\mathcal{H}_N}^2 = \|\kappa(\cdot,x)\|_{\mathcal{H}_N} - \|\mathcal{A}\kappa(\cdot,x)\|_{\overline{\mathcal{H}}_s} \tag{12}$$

Proof. Let us expand the second term in (10) as

$$\begin{aligned}&\|\kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x)\|_{\mathcal{H}_N}^2 \\ &= \langle \kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x), \kappa(\cdot,x) - \mathcal{P}\kappa(\cdot,x)\rangle_{\mathcal{H}_N} \\ &= \kappa(x,x) - 2\langle \kappa(\cdot,x), \mathcal{P}\kappa(\cdot,x)\rangle_{\mathcal{H}_N} + \langle \mathcal{P}\kappa(\cdot,x), \mathcal{P}\kappa(\cdot,x)\rangle_{\mathcal{H}_N} \\ &\overset{(a)}{=} \kappa(x,x) - 2\langle \kappa(\cdot,x), \mathcal{P}\kappa(\cdot,x)\rangle_{\mathcal{H}_N} + \langle \mathcal{A}\kappa(\cdot,x), \mathcal{A}\kappa(\cdot,x)\rangle_{\overline{\mathcal{H}}_s} \\ &\overset{(b)}{=} \kappa(x,x) - 2\langle \mathcal{A}\kappa(\cdot,x), \mathcal{A}\kappa(\cdot,x)\rangle_{\overline{\mathcal{H}}_s} + \langle \mathcal{A}\kappa(\cdot,x), \mathcal{A}\kappa(\cdot,x)\rangle_{\overline{\mathcal{H}}_s} \\ &= \kappa(x,x) - \langle \mathcal{A}\kappa(\cdot,x), \mathcal{A}\kappa(\cdot,x)\rangle_{\overline{\mathcal{H}}_s} \\ &= \|\kappa(\cdot,x)\|_{\mathcal{H}_N} - \|\mathcal{A}\kappa(\cdot,x)\|_{\overline{\mathcal{H}}_s} \end{aligned} \tag{13}$$

where, using $\langle f, \mathcal{A}^*\bar{\boldsymbol{f}}\rangle_{\mathcal{H}_N} = \langle \mathcal{A}f, \bar{\boldsymbol{f}}\rangle_{\overline{\mathcal{H}}_s}$ and $\mathcal{A}\mathcal{A}^* = \mathcal{I}$, the identity $\langle \mathcal{A}^*\mathcal{A}f, \mathcal{A}^*\mathcal{A}f\rangle_{\mathcal{H}_N} = \langle \mathcal{A}\mathcal{A}^*\mathcal{A}f, \mathcal{A}f\rangle_{\overline{\mathcal{H}}_s}$ has been applied in (a), and $\langle f, \mathcal{A}^*\mathcal{A}f\rangle_{\mathcal{H}_N} = \langle \mathcal{A}f, \mathcal{A}f\rangle_{\overline{\mathcal{H}}_s}$ in (b). □

Before continuing with the analysis, let us first introduce the following definition:

Nyström approximation. Let $\mathcal{B} \subseteq \mathbb{N}$ be a set indexing $|\mathcal{B}| = r \ll N$ columns of $\boldsymbol{K} \in \mathbb{R}^{N\times N}$. Defining $\boldsymbol{R} = [\boldsymbol{K}(i)]_{i\in\mathcal{B}}$ as the tall matrix formed by the indexed columns, its transpose $\boldsymbol{R}^T$, and $\boldsymbol{C}$ as the submatrix indexed by $\{(i,j)\}_{(i,j)\in\mathcal{B}\times\mathcal{B}}$, the Nyström approximation to $\boldsymbol{K}$ is

$$\boldsymbol{N} = \boldsymbol{R}\boldsymbol{C}^{\dagger}\boldsymbol{R}^T. \tag{14}$$

The Nyström approximation is a low-rank approximation that is used in place of the full-rank kernel matrix in kernel methods [15] to accelerate the inversion of the kernel matrix. While in this paper this matrix is not used for this purpose, it arises in the various analyses conducted from here on. Thus, although (10) addresses the per-sample error, Lemma 1 enables an upper bound formulated in terms of the Nyström approximation when considering the total reconstruction error across f. This bound constitutes the main statement in this section and is presented in the theorem below.

Theorem 1. *Let $\boldsymbol{f} = [f(x_1),\dots,f(x_N)]^T$, and $\boldsymbol{S}$ be an $s \times N$ binary sampling matrix with a non-zero element per row equal to 1 such that $\boldsymbol{S}\boldsymbol{f} := \mathcal{A}f$. Moreover, let $\boldsymbol{K}$ be the kernel matrix of $\mathcal{H}_N$, and $\boldsymbol{T} := \boldsymbol{K}\boldsymbol{S}^T\bar{\boldsymbol{K}}^{-1}\boldsymbol{S}\boldsymbol{K}$ with $\bar{\boldsymbol{K}} = \boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T$. Then, the absolute error across f, defined as $|f - \mathcal{P}f| = \sum_{m=1}^{N} |f(x_m) - \mathcal{P}f(x_m)|$, is bounded as*

$$|f - \mathcal{P}f| \leq \|f\|_{\mathcal{H}_N} \operatorname{Tr}(\boldsymbol{K} - \boldsymbol{T}) \tag{15}$$

Proof. Using (10), (13), and $\|\bar{\boldsymbol{f}}\|_{\overline{\mathcal{H}}_s} = \bar{\boldsymbol{f}}^T\bar{\boldsymbol{K}}^{-1}\bar{\boldsymbol{f}}$, it holds that

$$\begin{aligned}|f - \mathcal{P}f| &= \sum_{m=1}^{N} |f(x_m) - \mathcal{P}f(x_m)| \\ &\leq \|f\|_{\mathcal{H}_N} \sum_{m=1}^{N} \|\kappa(\cdot,x_m) - \mathcal{P}\kappa(\cdot,x_m)\|_{\mathcal{H}_N}^2 \\ &= \|f\|_{\mathcal{H}_N} \sum_{m=1}^{N} \left(\|\kappa(\cdot,x_m)\|_{\mathcal{H}_N} - \|\mathcal{A}\kappa(\cdot,x_m)\|_{\overline{\mathcal{H}}_s}\right) \\ &= \|f\|_{\mathcal{H}_N} \sum_{m=1}^{N} \left(\kappa(x_m,x_m) - \sum_{i=1}^{s}\sum_{j=1}^{s} \kappa(\bar{x}_i,x_m)(\bar{\boldsymbol{K}}^{-1})_{i,j}\kappa(\bar{x}_j,x_m)\right) \end{aligned} \tag{16}$$

$$\begin{aligned}&= \|f\|_{\mathcal{H}_N} \operatorname{Tr}(\boldsymbol{K} - \boldsymbol{K}\boldsymbol{S}^T\bar{\boldsymbol{K}}^{-1}\boldsymbol{S}\boldsymbol{K}) \\ &= \|f\|_{\mathcal{H}_N} \operatorname{Tr}(\boldsymbol{K} - \boldsymbol{T}). \end{aligned} \tag{17}$$

□

Matrix $\boldsymbol{T}$ constitutes a Nyström approximation as defined in (14), with the specificity that its rank, given by s, and the selection of columns is tied to the sampling matrix $\boldsymbol{S}$. Again, note that full-rank kernel matrix is not being replaced with an approximation, the Nyström approximation appears through the analysis conducted in Theorem 1. The result in (15) shows that the upper bound on the reconstruction error of f is proportional to the difference between $\boldsymbol{K}$ and $\boldsymbol{T}$ on its diagonal entries. Moreover, $\operatorname{Tr}(\boldsymbol{K} - \boldsymbol{T}) \geq 0$ since in (12) $\|\kappa(\cdot,x)\|_{\mathcal{H}_N} \geq \|\mathcal{A}\kappa(\cdot,x)\|_{\overline{\mathcal{H}}_s}$. Therefore, designing $\mathcal{A}$ to maximize $\sum_{n=1}^{N} \|\mathcal{A}\kappa(\cdot,x_n)\|_{\overline{\mathcal{H}}_s} = \operatorname{Tr}(\boldsymbol{T})$ minimizes the upper bound in Theorem 1.

3. Function sampling and reconstruction in kernel ridge regression

In KRR, the function to be recovered is a vector $\boldsymbol{f} = [f(x_1),\dots,f(x_N)]^T \in \mathcal{H}_N \subseteq \mathbb{R}^N$, where $\boldsymbol{f} = \boldsymbol{K}\boldsymbol{\alpha}$, $\boldsymbol{\alpha} = [\alpha_1,\dots,\alpha_N]^T$, and $\langle \boldsymbol{f}, \boldsymbol{f}'\rangle_{\mathcal{H}_N} = \boldsymbol{f}^T\boldsymbol{K}^{-1}\boldsymbol{f}'$ with $\boldsymbol{f}' \in \mathcal{H}_N$. Given the s-dimensional vector of noisy observations $\bar{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$, where $\boldsymbol{S}$ is an $s \times N$ binary sampling matrix with a single non-zero element per row, $\boldsymbol{y} = [y_1,\dots,y_N]^T$, $y_i = f(x_i) + w_i$ and w_i denotes noise, the problem is formulated as solving

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}\in\mathcal{H}_N}{\operatorname{argmin}} \|\bar{\boldsymbol{y}} - \boldsymbol{S}\boldsymbol{f}\|_2^2 + \mu\|\boldsymbol{f}\|_{\mathcal{H}_N} \tag{18}$$

or, equivalently,

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\operatorname{argmin}} \|\bar{\boldsymbol{y}} - \boldsymbol{S}\boldsymbol{K}\boldsymbol{\alpha}\|_2^2 + \mu\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha} \tag{19}$$

where $\mu > 0$ is a regularization parameter. The solution to (18) is

$$\hat{\boldsymbol{f}} = \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{y}. \tag{20}$$

As shown in Section 2, the sampling operator in (2) takes the form $\mathcal{A} := \boldsymbol{S}$, where $\boldsymbol{S}$ is an $s \times N$ binary sampling matrix with a single non-zero element per row such that $\boldsymbol{S}\boldsymbol{f} = \bar{\boldsymbol{f}}$. Its adjoint, derived following (6), is $\mathcal{A}^* := \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T)^{-1}$, and the projection operator is $\mathcal{P} = \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T)^{-1}\boldsymbol{S}$. Thus, one observes that the solution (20) is not obtained by applying the projection operator $\mathcal{P}$ to $\boldsymbol{y}$ unless $\mu = 0$. However, since (20) lies on the column span of $\boldsymbol{K}\boldsymbol{S}^T$, the result is a non-orthogonal projection of $\boldsymbol{y}$ onto $\mathcal{H}_s = \operatorname{span}\{\boldsymbol{K}\boldsymbol{S}^T\}$, i.e., the space spanned by the kernel functions evaluated at the sampled inputs. Still, a direct application of Theorem 1 is not possible unless $\mu = 0$ and $\boldsymbol{y}$ is noiseless.

To evaluate the performance of a specific sampling pattern, the risk function [16], which is equal to the mean squared error (MSE) except that it takes $\boldsymbol{S}$ as a parameter, will be used. Thus, the risk is

$$\begin{aligned} R_\mu(\boldsymbol{S}) &= \mathbb{E}_{\boldsymbol{w}}\{\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_2^2\} \\ &= \left\|(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S})\boldsymbol{f}\right\|_2^2 \\ &\quad + \mathbb{E}_{\boldsymbol{w}}\left\{\left\|\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{w}\right\|_2^2\right\} \end{aligned} \tag{21}$$

where $\boldsymbol{w} = [w_1, \dots, w_N]^T$. When $\mu = 0$, the risk becomes

$$\begin{aligned} R_0(\boldsymbol{S}) &= \left\|\boldsymbol{f} - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T)^{-1}\boldsymbol{S}\boldsymbol{f}\right\|_2^2 + \mathbb{E}_{\boldsymbol{w}}\left\{\left\|\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T)^{-1}\boldsymbol{S}\boldsymbol{w}\right\|_2^2\right\} \\ &= \|\boldsymbol{f} - \mathcal{P}\boldsymbol{f}\|_2^2 + \mathbb{E}_{\boldsymbol{w}}\left\{\|\mathcal{P}\boldsymbol{w}\|_2^2\right\}. \end{aligned} \tag{22}$$

Hence, the first term in (22), i.e., the squared bias, takes a similar form as (15) in Theorem 1, whereas the second term or variance is the norm of the projection of the vectorized noise onto $\mathcal{H}_s$.

In a noiseless situation, the variance term in (22) would disappear and therefore the overall error would depend uniquely on the accuracy of the Nyström approximation. With noise, there needs to be a balance between bias and variance in order to make the risk small, i.e., sampling the most important points in $\boldsymbol{f}$ while acquiring the least amount of noise. However, since the distribution of the noise is usually unknown, this strategy is not feasible. Assuming zero-mean Gaussian noise with variance ν^2, then $\mathbb{E}_{\boldsymbol{w}}\left\{\|\mathcal{P}\boldsymbol{w}\|_2^2\right\} = \nu^2\operatorname{Tr}(\mathcal{P}) = \nu^2 s$ since $\operatorname{Tr}(\mathcal{P}) = \operatorname{Tr}((\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T)^{-1}\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T) = s$. Adapting the procedure in the proof of Theorem 1 to the 2-norm, we have that

$$R_0(\boldsymbol{S}) = \|\boldsymbol{f} - \mathcal{P}\boldsymbol{f}\|_2^2 + \nu^2\operatorname{Tr}(\mathcal{P}) \le \|\boldsymbol{f}\|_2^2\operatorname{Tr}(\boldsymbol{K} - \boldsymbol{T}) + \nu^2 s$$

This expression shows that adding more samples reduces the approximation error to $\boldsymbol{K}$, and hence the bias, but it also increases the variance.

Since in KRR one typically uses $\mu > 0$ to reduce the impact of the noise and risk of overfitting, the passive sampling strategy is derived in the next section by analyzing $R_\mu(\boldsymbol{S})$ in (21).

4. Passive sampling for kernel ridge regression

The previous sections have shown that the upper bound on the recovery error increases with the trace of the difference between the kernel matrix and its Nyström approximation. Since this approximation is built from a subset of columns of the kernel matrix, it is crucial to choose the columns that best approximate the spectrum of the full matrix to keep the approximation accurate. Indeed, this is known as the column subset selection problem [17], also related to matrix sketching [18–20]. In regards to the Nyström approximation, while there exist deterministic [21] methods which order the columns according to a metric and then choose the top performers, most implementations opt for statistical approaches. In these, a sampling probability is assigned to each column and the required number of columns is sampled according to the resulting distribution. One example of such a technique is in [13,16], which measures column importance through the so-called leverage scores and provides theoretical guarantees on the approximation error. This section presents a passive sampling method based on kernel leverage scores, which is derived through an analysis of the estimation error in KRR.

4.1. Risk function analysis

Rewriting the risk function as shown in Appendix A, assuming Gaussian noise with variance ν^2, and using that $\|\boldsymbol{A}\boldsymbol{\alpha}\|_2 \le \|\boldsymbol{A}\|_2\|\boldsymbol{\alpha}\|_2$ for any matrix $\boldsymbol{A}$ of appropriate size, the risk is upper bounded as

$$\begin{aligned} R_\mu(\boldsymbol{S}) &= \left\|(\boldsymbol{K} - \tilde{\boldsymbol{T}})\boldsymbol{\alpha}\right\|_2^2 + \mathbb{E}_{\boldsymbol{w}}\left\{\frac{1}{\mu^2}\left\|(\boldsymbol{K} - \tilde{\boldsymbol{T}})\boldsymbol{S}^T\tilde{\boldsymbol{w}}\right\|_2^2\right\} \\ &\le \left\|\boldsymbol{K} - \tilde{\boldsymbol{T}}\right\|_2^2\left(\|\boldsymbol{\alpha}\|_2^2 + \frac{\nu^2 s}{\mu^2}\right) \end{aligned} \tag{23}$$

where $\tilde{\boldsymbol{T}} := \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{K}$ is a regularized Nyström approximation.

Equation (23) shows that the spectral norm of the difference between the kernel matrix and its regularized Nyström approximation, which depends on μ, also determines the risk bound for KRR. Therefore, it is well-grounded that one may pick the training inputs in $\mathcal{X}$ by choosing the associated columns in $\boldsymbol{K}$ that best build the regularized Nyström approximation. This section frames the sampling of the entries in $\boldsymbol{y}$ as designing the sampling matrix $\boldsymbol{S}$ that minimizes the error $\boldsymbol{K} - \tilde{\boldsymbol{T}}$. Through analyzing the risk function from a probabilistic standpoint, $\boldsymbol{S}$ is cast as a weighted sampling matrix with weights set according to the leverage scores of the columns of $\boldsymbol{K}$. Since this process only involves the kernel matrix, it is a passive sampling approach. Once $\boldsymbol{S}$ is obtained, the corresponding samples are obtained as $\bar{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and KRR can then be applied to recover $\boldsymbol{f}$.

Let us first introduce the following lemma proven in Appendix B:

Lemma 2. *The difference $\boldsymbol{K} - \tilde{\boldsymbol{T}}$ can be written as*

$$\boldsymbol{K} - \tilde{\boldsymbol{T}} = \mu\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{P})^{-1}(\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T \tag{24}$$

where we have used the eigendecomposition $\boldsymbol{K} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T$ and

$$\boldsymbol{P} := \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-1} - (\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-\frac{1}{2}}. \tag{25}$$

Defining

$$\boldsymbol{V} = (\boldsymbol{\Sigma} + \mu\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T \tag{26}$$

and knowing that the product of two diagonal matrices is commutative, substituting into (25) yields

$$\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{V}^T \tag{27}$$

Therefore, $\boldsymbol{P}$ is the difference between $\boldsymbol{V}\boldsymbol{V}^T$ and an approximation built from a subset of the columns in $\boldsymbol{V}$, and it is the only element in (24) depending on the sampling distribution through $\boldsymbol{S}$. Thus, if $\boldsymbol{S}$ is designed to minimize the norm of $\boldsymbol{P}$ so that in (24) $\|(\boldsymbol{I} - \boldsymbol{P})^{-1}\|_2$ is approximately minimized, a reduction on the difference $\boldsymbol{K} - \tilde{\boldsymbol{T}}$ in (24) and also on the bound on $R_\mu(\boldsymbol{S})$ in (23) will be achieved.

4.2. Sampling based on leverage scores

Let θ be an ordered s-tuple containing indices drawn with replacement from $\{1,\dots,N\}$ with the probability distribution $\Pr(i) = p_i$. Here, p_i denotes the probability of sampling the ith column in $\boldsymbol{K}$, and θ contains the indices of the sampled columns in non-descending order. Moreover, assume a weighted $\boldsymbol{S} \in \mathbb{R}^{s\times N}$ with entries

$$\boldsymbol{S}_{i,j} = \begin{cases} \frac{1}{\sqrt{sp_j}} & \text{if } \theta(i) = j \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

for $i = 1,\dots,s$ and $j = 1,\dots,N$. Note that, since the indices in θ are drawn with replacement, there might be repeated rows in $\boldsymbol{S}$ sampling the same item. Moreover, a weighted $\boldsymbol{S}$ is necessary in order to satisfy certain theoretical guarantees.

In [22] it is shown that the optimal probability distribution to minimize $\mathbb{E}\{\|\boldsymbol{P}\|_{\mathrm{F}}^2\}$ in (27) is

$$p_i = \frac{\|\boldsymbol{v}_i\|_2^2}{\|\boldsymbol{V}\|_{\mathrm{F}}^2} \tag{29}$$

with $\boldsymbol{v}_i$ denoting the ith column of $\boldsymbol{V}$ in (26), and

$$\|\boldsymbol{v}_i\|_2^2 = \sum_{j=1}^{N} \frac{\sigma_j}{\sigma_j+\mu} \boldsymbol{q}_j^2(i) \tag{30}$$

where $\sigma_i = (\boldsymbol{\Sigma})_{ii}$, and $\boldsymbol{q}_j$ is the jth column of $\boldsymbol{Q}$. Thus, the distribution obtained with (29) for $i = 1,\dots,N$ can be expected to reduce $\|\boldsymbol{K} - \tilde{\boldsymbol{T}}\|$ in the risk upper bound (23). Still, note that this distribution is obtained for a specific value of μ since p_i depends on μ through $\boldsymbol{V}$.

The KRR regularization parameter μ is usually chosen experimentally via cross-validation to minimize the prediction error on a testing dataset. However, as it will be shown in the numerical results, it is possible to further reduce the prediction error by fine-tuning the sampling distribution. For this reason, we decouple the optimal distribution from the regularization parameter in KRR by redefining the ith leverage score as

$$l_i = \sum_{j=1}^{N} \frac{\sigma_j}{\sigma_j+\gamma} \boldsymbol{q}_j^2(i) \tag{31}$$

where $\gamma > 0$ is now a tunable parameter different from the parameter μ. Alternatively, (31) can be obtained as $l_i = (\boldsymbol{K}(\boldsymbol{K}+\gamma\boldsymbol{I})^{-1})_{i,i}$. Then, the chosen distribution is

$$p_i = \frac{l_i}{\sum_{j=1}^{N} l_j}. \tag{32}$$

Thus, replacing μ in (30) with γ to obtain (31) enables separate optimization of the regularization term in the regression problem and the probability distribution (32). We refer to the use of this distribution to select the training set as *leverage score sampling*.

The quantity l_i is the so-called regularized leverage score[1] of the ith column of $\boldsymbol{K}$, a concept that often appears in the context of random matrix approximations. For instance, [13,16] derive error bounds on the Nyström approximation when the columns are sampled according to their leverage scores. Similarly, [23] relies on the leverage scores to sample a subset of columns of the kernel matrix and reduce the computational cost of KRR. In [24], leverage scores are used to assess the approximation of the kernel matrix via random Fourier features, whereas [18] uses them to obtain accurate subspace embeddings. In contrast to these related works, our focus is not on obtaining the best possible kernel matrix approximation to reduce computational cost in KRR. We set out to derive an optimal approach to select the samples in the training set, and the analysis of the risk function has led us to the leverage scores (31) as a good approach.

Since the ith leverage score is a weighted average of the ith row of the eigenvector matrix $\boldsymbol{Q}$, a high value indicates that the ith input stands out with respect to the other elements. Moreover, γ determines how much weight is given to the notable points, with $\gamma = 0$ inducing a uniform probability distribution. Regarding computational cost, it is dominated by the inversion of an $N \times N$ matrix, which could be reducer by replacing the matrix to be inverted with a low-rank approximation. This is for instance proposed in [16] where, given an $N \times d$ binary sampling matrix $\boldsymbol{D}$, an approximate leverage score is obtained as

$$\tilde{l}_i = \frac{1}{\gamma}\left(\boldsymbol{K} - \boldsymbol{K}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{K}\boldsymbol{D} + \gamma\boldsymbol{I})^{-1}\boldsymbol{S}^T\boldsymbol{K}\right). \tag{33}$$

If the columns of $\boldsymbol{K}$ are sampled according to p_i in (32), the risk for a given $\boldsymbol{S}$ can be bounded in relation to the risk of the full problem. To do so, let us first introduce the following theorem adapted from [13] to bound the spectral norm of $\boldsymbol{P}$ in (27):

Theorem 2. *[13] Let $\boldsymbol{V} \in N \times N$, and $\boldsymbol{S}$ be an $s \times N$ sparse weighted sampling matrix with $\boldsymbol{S}_{i,j} = \frac{1}{\sqrt{sp_j}}$ if $\theta(i) = j$ $\forall i = 1,\dots,s$, $j \in \{1,\dots,N\}$, where p_j denotes the probability of choosing the jth column in $\boldsymbol{K}$. Then,*

$$\Pr\left(\left\|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{V}^T\right\|_2 \geq t\right) \leq N\exp\left(\frac{-st^2/2}{\|\boldsymbol{V}\|_{\mathrm{F}}^2 + t/3}\right). \tag{34}$$

Using Theorem 2, $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{V}^T \preceq t\boldsymbol{I}$ with probability greater than $1-\delta$ if $s \geq (\frac{2d_{eff}}{t^2} + \frac{2}{3t})\log\frac{N}{\delta}$, where d_{eff} is the so-called effective dimension defined as $d_{eff} = \sum_{i=1}^{N} l_i = \|\boldsymbol{V}\|_{\mathrm{F}}^2$, and $\delta \in (0,1]$; this condition on s is obtained by equating the right-hand side of (34) to δ. Then, we have that $(\boldsymbol{I}-\boldsymbol{P})^{-1} \preceq \frac{1}{1-t}\boldsymbol{I}$ and, therefore, the eigenvalues of $\boldsymbol{K} - \tilde{\boldsymbol{T}}$ in (24) are bounded as

$$\boldsymbol{K} - \tilde{\boldsymbol{T}} \preceq \frac{\mu}{1-t}\boldsymbol{I} \tag{35}$$

with probability at least $1-\delta$. This inequality is obtained by applying to (24) the property that the maximum eigenvalue of a matrix multiplication is smaller or equal to the multiplication of the maximum eigenvalue of each matrix [10]. Using (35), the risk is written in relation to the risk of the fully observed problem, i.e., $\boldsymbol{S} = \boldsymbol{I}$, as Theorem 3 below shows with proof in Appendix C.

Theorem 3. *Let $\boldsymbol{f} = \boldsymbol{K}\boldsymbol{\alpha}$ be the function vector to be recovered, $R_\mu(\cdot)$ the KRR risk function in (21), and $\boldsymbol{S}$ a $s \times N$ sparse weighted sampling matrix with $\boldsymbol{S}_{i,j} = \frac{1}{\sqrt{sp_j}}$ if $\theta(i) = j$ $\forall i = 1,\dots,s$, $j \in \{1,\dots,N\}$, where p_j denotes the probability of choosing the jth column in $\boldsymbol{K}$. Provided that $s \geq (\frac{2d_{eff}}{t^2} + \frac{2}{3t})\log\frac{N}{\delta}$, it holds with probability greater than $1-\delta$ for $\delta \in (0,1]$ that the risk is upper bounded as*

$$R_\mu(\boldsymbol{S}) \leq R_\mu(\boldsymbol{I}) + \frac{\mu t}{1-t}\|\boldsymbol{\alpha}\|_2^2 + \frac{\nu^2}{(1-t)^2}\mathrm{Tr}(\boldsymbol{S}^T\boldsymbol{S} - \boldsymbol{I}). \tag{36}$$

Theorem 3 indicates that the excess risk when using $s < N$ samples increases with the spectral norm of $\boldsymbol{P}$ assumed smaller than t, with the second term in (36) being related to the excess bias, and the third to the excess variance; see (52) and (55) in Appendix C. As expected, increasing s reduces the bias through inducing a smaller t, whereas a larger number of samples may result in higher variance due to the additional noise.

While it does not have the same theoretical guarantees as the probabilistic approach, a greedy approach can also be derived using the leverage scores. Greedy methods are concerned with maximizing optimality in the stage at which they are executed, and

[1] Hereafter the words leverage score refer to the regularized leverage score.

do not consider the optimality of the final solution. Applied to the passive sampling paradigm we have developed, this means simply taking the samples with the highest associated leverage score. Defining $\mathcal{L}$ as the set containing the s leverage scores with the highest value, the s-tuple of observation indices for *greedy leverage sampling* is denoted by

$$\theta_G = (i)_{l_i \in \mathcal{L}}. \tag{37}$$

Remark 1. Sampling based on leverage scores is applicable to the selection of inducing points in Gaussian processes. Assuming that f is a Gaussian process (GP) such that $\Pr(f|\mathcal{X}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K})$, the distribution of f conditioned on the observations, inputs and variance v^2 is

$$\Pr(f|\bar{\boldsymbol{m}}, \mathcal{X}, v^2) \sim \mathcal{N}\Big[\boldsymbol{K}\boldsymbol{S}^T(\bar{\boldsymbol{K}} + v^2\boldsymbol{I})^{-1}\bar{\boldsymbol{y}}, \boldsymbol{K} - \boldsymbol{K}\boldsymbol{S}^T(\bar{\boldsymbol{K}} + v^2\boldsymbol{I})^{-1}\boldsymbol{K}\boldsymbol{S}\Big]. \tag{38}$$

We observe in (38) that the mean is equal to a KRR with $\mu = v^2$, and that the covariance matrix is the difference between $\boldsymbol{K}$ and its Nyström approximation. Hence, this suggests that an accurate approximation through the design of $\boldsymbol{S}$ should yield a lower variance on the posterior distribution of f. Such a design can be produced by means of the leverage scores as presented in this section, which in turn induces a sampling distribution through $\bar{\boldsymbol{y}}$. Hence, to obtain either a leverage score sampling or greedy leverage sampling scheme for GPs, one just needs to calculate the leverage scores as in (31) for a specific γ and follow the indicated steps for each of the two methods.

5. Numerical tests

This section presents numerical tests comparing uniform sampling with $p_i = \frac{1}{N}$ to leverage score sampling (32) and greedy leverage score sampling in (37) for different values of s, for both KRR and GP regression. Each combination of sampling method and algorithm is labeled as UKRR, LKRR, GKRR, UGP, LGP and GGP, where U, L and G stand for uniform, leverage and greedy, respectively. The optimal hyperparameters μ and γ are found via grid search for each s. The search is performed randomly: intervals $[\mu_{\min}, \mu_{\max}]$ and $[\gamma_{\min}, \gamma_{\max}]$ are set for μ and γ respectively, and 100 tuples of values within the intervals are selected at random. For the GP algorithm, the noise variance v^2 is estimated through the log-likelihood. The s labeled samples form the training set, and the RMSE measures the estimation error as

$$\text{RMSE} = \sqrt{\frac{1}{N_r}\sum_{i=1}^{N_r}\left\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}\right\|_{\text{F}}^2} \tag{39}$$

where $N_r = 20$ is the number of realizations, $\hat{\boldsymbol{f}}_i$ is the estimation at the ith realization, and a new training set is acquired at each realization. For KRR, $\hat{\boldsymbol{f}}_i$ is given by (20), whereas for a GP it is given by the posterior mean in (38) with v obtained via maximum-likelihood estimation from the observed samples.

The Boston housing dataset[2] is the first to be evaluated. The input set $\mathcal{X}$ is comprised of 506 feature vectors detailing the characteristics of 506 houses in Boston such as size or number of rooms. The variable to be predicted, i.e., $\boldsymbol{f}$, is the price of each house. Here, the sampling algorithm selects a training set s of houses with their features, requests their prices, and estimates the price of the remaining ones. The used kernel is the Gaussian kernel applied on the feature vectors; its expression is $\boldsymbol{K}_{i,j} = \exp(\|x_i - x_j\|_2^2/\xi)$, where $\xi > 0$. Fig. 1 shows the RMSE for different values of s and the three sampling strategies. The figure shows that leverage score sampling attains a smaller error than uniform sampling for both KRR and GP regression. While the gains are small, they are in line with those obtained by other methods [3,13]. Nevertheless, for this particular dataset greedy sampling provides better performance for both algorithms.

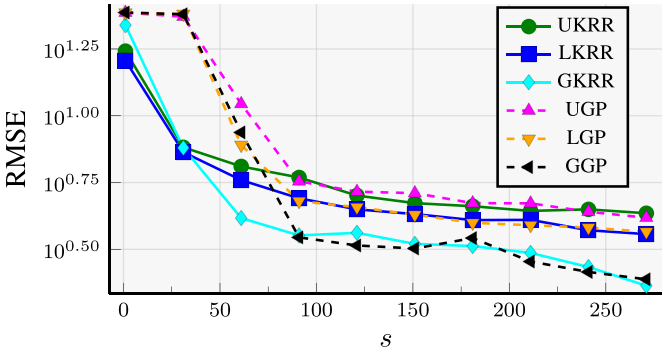

Fig. 1. NMSE vs. s for the Boston housing data.

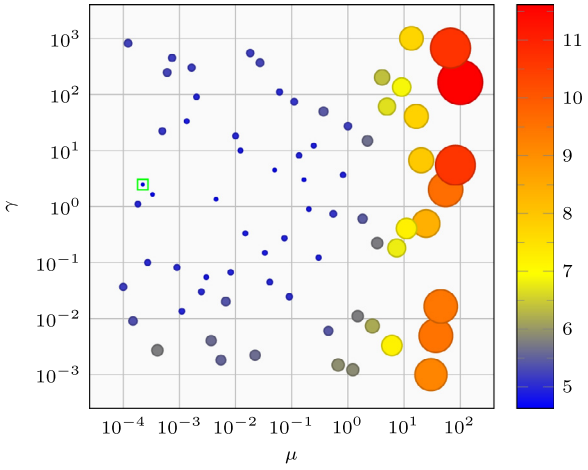

Fig. 2. Grid search for the Boston housing dataset for $s = 91$. The minimum is highlighted in a green square, centered at $\mu^* = 2.2 \cdot 10^{-4}$, $\gamma^* = 2.46$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2 depicts the result of the grid search for the Boston housing dataset for the LKRR method with $s = 91$. We observe that, out of all the evaluated points, the combination of μ and γ minimizing the error is $\mu^* = 2.2 \cdot 10^{-4}$, $\gamma^* = 2.46$. This plot corroborates that decoupling the hyperparameter for the design of the sampling distribution from the one in the KRR regularization yields a smaller error.

To gain further insight into the impact of γ, Fig. 3 shows the sampling probability distribution for several values of γ. We observe that setting a very small value or one large enough results in a more uniform distribution, although for $\gamma = 10$ the largest peaks still remain visible. On the other hand, an in-between value enhances the variability between probabilities. Thus, in the specific case in Fig. 2 with $s = 91$, the optimal γ is the one that maintains the largest sampling probabilities while promoting uniformity among the rest.

Fig. 4 shows the results for the Mushroom dataset. The vector to be recovered is of length 5,000 and the kernel is obtained from the Pearson correlation matrix of the features. Here, the feature vectors contain details about each mushroom such as shape or size, and the objective is to determine whether a mushroom is edible or poisonous. For KRR, both greedy and leverage score sampling show a much lower error than uniform sampling. On the other hand, in GP regression greedy sampling results in worse per-

[2] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html.

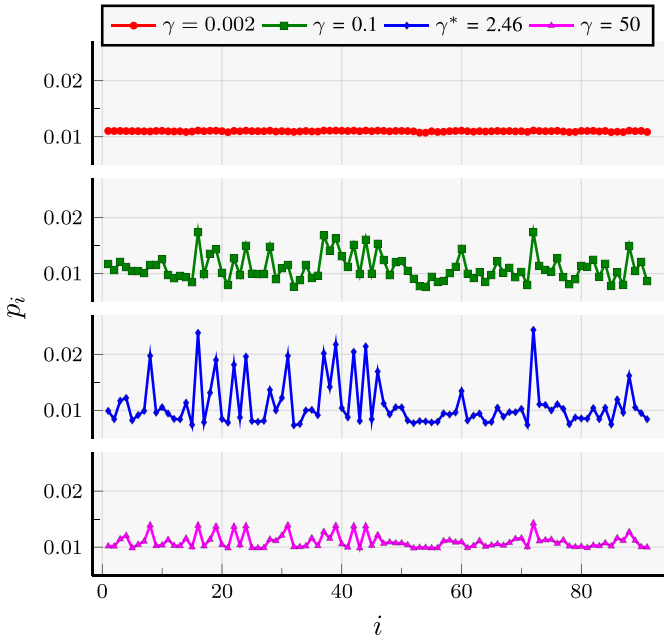

Fig. 3. Leverage scores and probabilities for the Boston housing dataset for different values of γ.

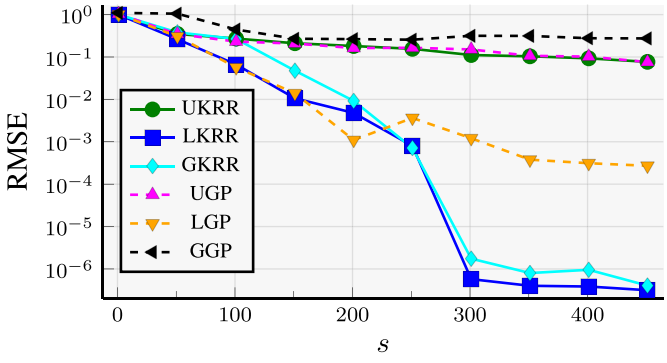

Fig. 4. RMSE vs. s for the Mushroom data.

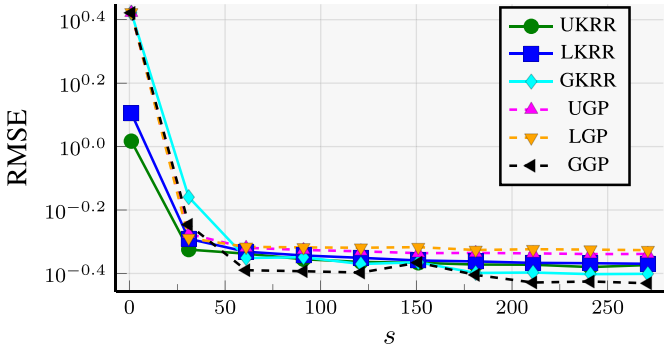

Fig. 5. RMSE vs. s for the Labor data.

formance when compared to uniform sampling. Note that, since this is a binary classification problem with classes $[2, -1]$, the NMSE evaluates the difference between the regression result and the actual numerical value of the class before applying any decision rule.

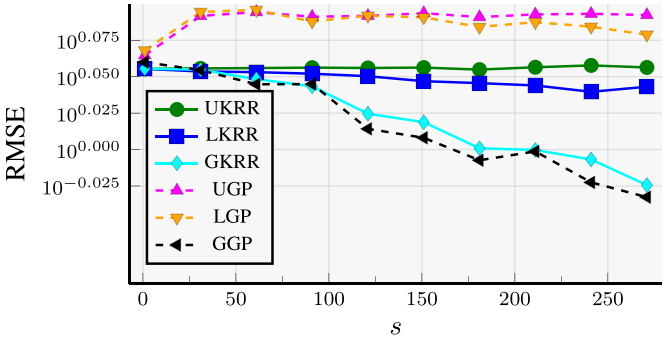

Fig. 6. RMSE vs. s for the Stocks data.

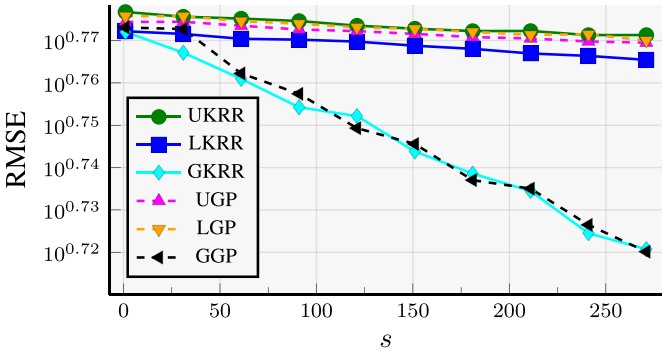

Fig. 7. RMSE vs. s for the Wine data.

Figs. 5, 6 and 7 show the test results for the Labor[3], Stocks[4] and Wine[5] datasets, respectively, using the Gaussian kernel. Overall, we observe that the passive sampling approaches are able to reduce the error with respect to uniform sampling, with the reduction being more noticeable for greedy sampling in the Stocks and Wine datasets.

We can conclude that, for KRR, both the random leverage sampling and the deterministic greedy sampling result in smaller error than uniform sampling for the test data. However, greedy sampling is more vulnerable to a badly designed kernel and outliers as it always chooses the most distinct points; this can be mitigated in leverage sampling by setting a more uniform distribution. Moreover, it should be noted that the reported results show the average error over several realizations for leverage sampling whereas for greedy sampling they correspond to a single realization since the sampled set is deterministic.

5.1. Computational cost evaluation

In order to reduce the computational cost, one may resort to approximate leverage scores. Fig. 8 shows the time required to complete the passive sampling using the approximate scores in (33) plus the KRR (18) for the Mushroom data. We observe that the cost is reduced in proportion to the factor $\frac{d}{N}$. Moreover, since the cost to calculate the leverage scores is constant for different s as it only depends on $\frac{d}{N}$, the increase in time with s is due to the KRR step. Fig. 9 shows the resulting probability distribution for the different factor values and first 100 indices. Finally, Fig. 10 compares the error for KRR using exact leverage scores, and approximate leverage scores with $\frac{d}{N} = 0.8$. We observe that at this value

[3] https://rdrr.io/rforge/Ecdat/man/Mroz.html.
[4] https://rdrr.io/cran/ISLR/man/Smarket.html.
[5] https://archive.ics.uci.edu/ml/datasets/wine.

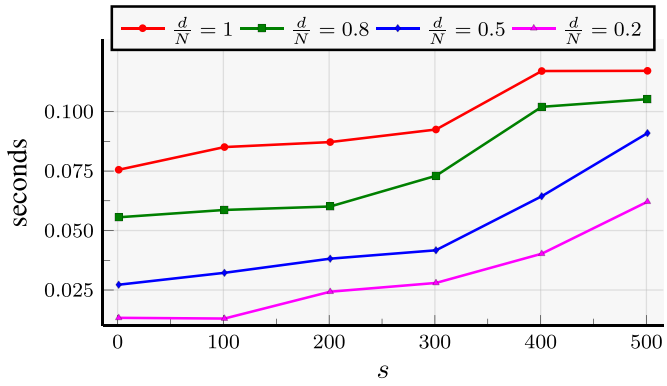

Fig. 8. Runtime for different values of $\frac{d}{N}$ on the Mushroom data.

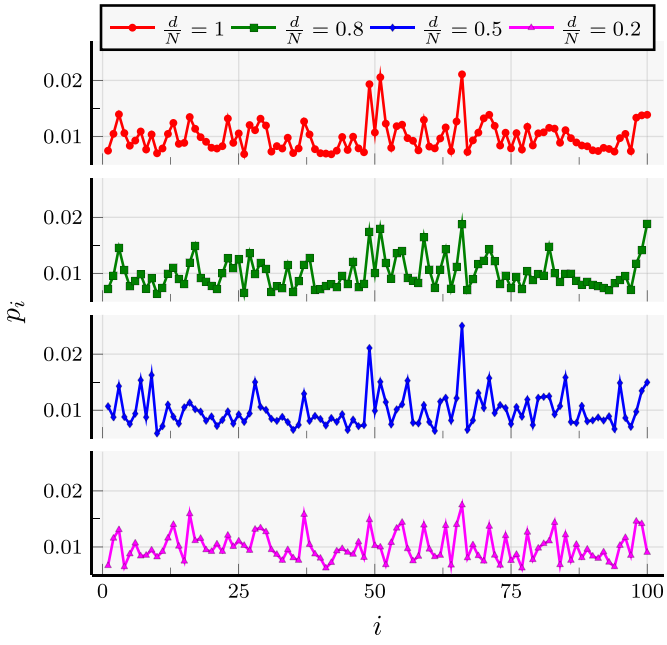

Fig. 9. Sampling probabilities of the first 100 points in the Mushroom data for different $\frac{d}{N}$.

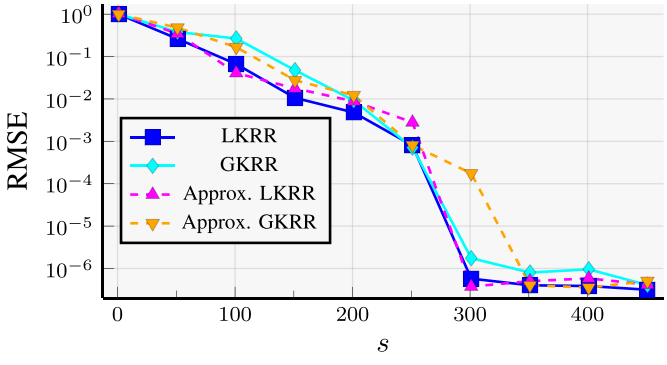

Fig. 10. RMSE vs. s for the Mushroom data using exact leverage scores, and approximate leverage scores with $\frac{d}{N} = 0.8$.

the performance is not significantly degraded, while the computational cost is reduced.

6. Conclusions

This paper has studied how the prior information embedded in the kernel functions spanning an RKHS can be used to determine which inputs are the most important to label in order to recover the complete function. First, through functional analysis, it has been shown that the recovery error of a sampled function in an RKHS is directly tied to the Nyström approximation to the kernel matrix. Hence, in the noiseless case, the error can be minimized by finding the best possible Nyström approximation and labeling the inputs corresponding to the chosen columns. This sampling approach is passive since it only involves knowledge of the input space and kernel matrix, which presents benefits over active sampling schemes which require online operation and are more vulnerable to noise. Given the theoretical background, the Nyström-based sampling approach has been applied to KRR and GP regression for the picking of the training set. In this, the weighted sampling matrix is designed according to the leverage scores of the kernel matrix, which measure the importance of each column in achieving a good Nyström approximation. Numerical tests have shown that the proposed approaches work well for the recovery of sampled vectors when compared to uniform sampling.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Prof. Daniel Romero from University of Adger, and Prof. Antonio G. Marqués from Universidad Rey Juan Carlos for their valuable insights.

Appendix A. Risk function for KRR

For KRR, the MSE is

$$MSE := \mathbb{E}_{\boldsymbol{w}}\left\{\|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_2^2\right\} = \mathbb{E}_{\boldsymbol{w}}\left\{\left\|\boldsymbol{f} - \boldsymbol{K}\hat{\boldsymbol{\alpha}}\right\|_2^2\right\}. \tag{39}$$

Plugging the estimator from (20) into (39) yields

$$\begin{aligned} MSE &= \mathbb{E}_{\boldsymbol{w}}\left\{\left\|\boldsymbol{f} - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}(\boldsymbol{S}\boldsymbol{f} + \bar{\boldsymbol{w}})\right\|_2^2\right\} \\ &= \left\|(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S})\boldsymbol{f}\right\|_2^2 \\ &\quad + \mathbb{E}_{\boldsymbol{w}}\left\{\left\|\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\bar{\boldsymbol{w}}\right\|_2^2\right\} \end{aligned} \tag{40}$$

where $\bar{\boldsymbol{w}} = \boldsymbol{S}\boldsymbol{w}$, and we have used that $\mathbb{E}\{\boldsymbol{w}\} = \boldsymbol{0}$. Further, the first and second terms in (40) are the squared bias and variance of the KRR estimator, respectively. If we substitute $\boldsymbol{f} = \boldsymbol{K}\boldsymbol{\alpha}$ into the first term of (40), we obtain the squared bias as

$$\begin{aligned} b^2 &= \left\|(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S})\boldsymbol{K}\boldsymbol{\alpha}\right\|_2^2 \\ &= \left\|(\boldsymbol{K} - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{K})\boldsymbol{\alpha}\right\|_2^2 \\ &= \left\|(\boldsymbol{K} - \tilde{\boldsymbol{T}})\boldsymbol{\alpha}\right\|_2^2 \end{aligned} \tag{41}$$

where $\tilde{\boldsymbol{T}} := \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S})\boldsymbol{K}$ is the regularized Nyström approximation of $\boldsymbol{K}$. On the other hand, the variance term is

$$\begin{aligned} var &= \mathbb{E}_{\boldsymbol{w}}\left\{\left\|\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\bar{\boldsymbol{w}}\right\|_2^2\right\} \\ &= \mathbb{E}_{\boldsymbol{w}}\left\{\frac{1}{\mu^2}\left|\left|\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}(\mu\boldsymbol{I} + \boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T - \boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T)\bar{\boldsymbol{w}}\right|\right|_2^2\right\} \\ &= \mathbb{E}_{\boldsymbol{w}}\left\{\frac{1}{\mu^2}\left\|\boldsymbol{K}\boldsymbol{S}^T - \boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T + \mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T\bar{\boldsymbol{w}}\right\|_2^2\right\} \end{aligned}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left\{\frac{1}{\mu^2}\left\|(\boldsymbol{K}-\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T+\mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{K})\boldsymbol{S}^T\bar{\boldsymbol{w}}\right\|_2^2\right\}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left\{\frac{1}{\mu^2}\left\|(\boldsymbol{K}-\tilde{\boldsymbol{T}})\boldsymbol{S}^T\bar{\boldsymbol{w}}\right\|_2^2\right\}. \quad (42)$$

Adding the two terms in (41) and (42), and assuming a fixed $\boldsymbol{S}$, we obtain the risk in (23).

Appendix B. Proof of Lemma 2

With the eigendecomposition $\boldsymbol{K}=\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^T$, we can write

$$\boldsymbol{K}-\tilde{\boldsymbol{T}} = \boldsymbol{K}-\boldsymbol{K}\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S}^T+\mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{K}$$

$$= \boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\left[\boldsymbol{I}-\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T+\mu\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\right]\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T. \quad (43)$$

Applying the Matrix Inversion Lemma to the matrix inside the square brackets of (43), we arrive at

$$\boldsymbol{I}-\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T\left(\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T+\mu\boldsymbol{I}\right)^{-1}\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}$$

$$= \left(\boldsymbol{I}+\frac{1}{\mu}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}. \quad (44)$$

That in turn implies

$$\boldsymbol{K}-\tilde{\boldsymbol{T}} = \mu\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\mu\boldsymbol{I}+\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T$$

$$= \mu\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\boldsymbol{\Sigma}+\mu\boldsymbol{I}-\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T$$

$$= \mu\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\Big[(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{\frac{1}{2}}\Big(\boldsymbol{I}-(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{-\frac{1}{2}}$$

$$+(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{-\frac{1}{2}}\Big)$$

$$(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{\frac{1}{2}}\Big]^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T$$

$$= \mu\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{I}-\boldsymbol{P})^{-1}(\boldsymbol{\Sigma}+\mu\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}^T \quad (45)$$

with $\boldsymbol{P}$ given in (25).

Appendix C. Proof of Theorem 3

Let us first introduce the following lemma:

Lemma 3. *Let* $\check{\boldsymbol{K}}=\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I})^{-1}\boldsymbol{K}$. *Then,*

$$\boldsymbol{K}-\check{\boldsymbol{K}} = \mu\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I}) \quad (46)$$

Proof. Let

$$\boldsymbol{K}-\check{\boldsymbol{K}} = \boldsymbol{K}(\boldsymbol{I}-\frac{1}{\mu}(\frac{1}{\mu}\boldsymbol{K}+\boldsymbol{I})^{-1}\boldsymbol{K}). \quad (47)$$

Using the Matrix Inversion Lemma,

$$\boldsymbol{A}^{-1}+\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})\boldsymbol{C}\boldsymbol{A} = (\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \quad (48)$$

with $\boldsymbol{A}=\boldsymbol{B}=\boldsymbol{D}=\boldsymbol{I}$ and $\boldsymbol{C}=-\frac{1}{\mu}\boldsymbol{K}$, we have that

$$\boldsymbol{K}(\boldsymbol{I}-\frac{1}{\mu}(\frac{1}{\mu}\boldsymbol{K}+\boldsymbol{I})^{-1}\boldsymbol{K}) = \boldsymbol{K}(\frac{1}{\mu}\boldsymbol{K}+I)^{-1} = \mu\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I})^{-1}. \quad (49)$$

□

Knowing that $R_\mu(\boldsymbol{S})=R_\mu(\boldsymbol{I})+(R_\mu(\boldsymbol{S})-R_\mu(\boldsymbol{I}))$, we first write the bias in terms of the one for the full problem, namely $b_{\boldsymbol{I}}$, as

$$b^2 = \|(\boldsymbol{K}-\tilde{\boldsymbol{T}})\boldsymbol{\alpha}\|_2^2 \le \|(\boldsymbol{K}-\check{\boldsymbol{K}})\boldsymbol{\alpha}\|_2^2+\|(\check{\boldsymbol{K}}-\tilde{\boldsymbol{T}})\boldsymbol{\alpha}\|_2^2$$

$$= b_{\boldsymbol{I}}^2+\|(\check{\boldsymbol{K}}-\tilde{\boldsymbol{T}})\boldsymbol{\alpha}\|_2^2 \quad (50)$$

Provided that $s \ge (\frac{2d_{eff}}{t^2}+\frac{2}{3t})\log\frac{N}{\delta}$, from (35) we know that $\boldsymbol{K}-\tilde{\boldsymbol{T}} \preceq \frac{\mu}{1-t}\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I})^{-1}$ with probability $1-\delta$. Hereafter assume that any equation involving t is satisfied with probability $1-\delta$. Thus, to bound the last term in (50) we use Lemma 3 to show that

$$\check{\boldsymbol{K}}-\tilde{\boldsymbol{T}} = (\boldsymbol{K}-\tilde{\boldsymbol{T}})-(\boldsymbol{K}-\check{\boldsymbol{K}})$$

$$\preceq \frac{\mu}{1-t}\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I})^{-1}-\mu\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I})^{-1}$$

$$= \frac{\mu t}{1-t}\boldsymbol{K}(\boldsymbol{K}+\mu\boldsymbol{I})^{-1} \preceq \frac{\mu t}{1-t} \quad (51)$$

Applying the identity $\|(\check{\boldsymbol{K}}-\tilde{\boldsymbol{T}})\boldsymbol{\alpha}\|_2^2 \le \lambda_{\max}(\check{\boldsymbol{K}}-\tilde{\boldsymbol{T}})\|\boldsymbol{\alpha}\|_2^2$ yields

$$b^2 = b_{\boldsymbol{I}}^2+\frac{\mu t}{1-t}\|\boldsymbol{\alpha}\|_2^2. \quad (52)$$

Next, we proceed to calculate an upper bound to the difference between the variance of the sampled problem and that of the full problem, which is denoted as $var_{\boldsymbol{K}}$. Assuming Gaussian noise with variance ν^2 in (42), let $var=\frac{\nu^2}{\mu^2}\mathrm{Tr}((\boldsymbol{K}-\tilde{\boldsymbol{T}})^2\boldsymbol{S}^T\boldsymbol{S})$ and $var_{\boldsymbol{I}}=\frac{\nu^2}{\mu^2}\mathrm{Tr}((\boldsymbol{K}-\tilde{\boldsymbol{T}})^2)$. Then,

$$var-var_{\boldsymbol{I}} = \frac{\nu^2}{\mu^2}\mathrm{Tr}((\boldsymbol{K}-\tilde{\boldsymbol{T}})^2\boldsymbol{S}^T\boldsymbol{S})-\frac{\nu^2}{\mu^2}\mathrm{Tr}((\boldsymbol{K}-\tilde{\boldsymbol{T}})^2)$$

$$= \frac{\nu^2}{\mu^2}\mathrm{Tr}((\boldsymbol{K}-\tilde{\boldsymbol{T}})^2(\boldsymbol{S}^T\boldsymbol{S}-\boldsymbol{I})) \quad (53)$$

Applying the property $\boldsymbol{K}-\tilde{\boldsymbol{T}} \preceq \frac{\mu}{1-t}$ from (35), this difference can be upper bounded as

$$\frac{\nu^2}{\mu^2}\mathrm{Tr}((\boldsymbol{K}-\tilde{\boldsymbol{T}})^2(\boldsymbol{S}^T\boldsymbol{S}-\boldsymbol{I})) \le \frac{\nu^2}{(1-t)^2}\mathrm{Tr}(\boldsymbol{S}^T\boldsymbol{S}-\boldsymbol{I}) \quad (54)$$

Thus,

$$var \le var_{\boldsymbol{I}}+\frac{\nu^2}{(1-t)^2}\mathrm{Tr}(\boldsymbol{S}^T\boldsymbol{S}-\boldsymbol{I}). \quad (55)$$

Adding together (52) and (55) yields the upper bound in the theorem.

References

[1] C.E. Rasmussen, Gaussian processes in machine learning, in: Summer School on Machine Learning, Springer, 2003, pp. 63–71.

[2] F. Pérez-Cruz, S. Van Vaerenbergh, J.J. Murillo-Fuentes, M. Lázaro-Gredilla, I. Santamaria, Gaussian processes for nonlinear signal processing: an overview of recent advances, IEEE Signal Process. Mag. 30 (4) (2013) 40–50.

[3] H. Yu, S. Kim, Passive sampling for regression, in: 2010 IEEE International Conference on Data Mining, 2010, pp. 1151–1156, doi:10.1109/ICDM.2010.9.

[4] Q. Gu, J. Han, Towards active learning on graphs: an error bound minimization approach, in: IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 882–887.

[5] C. Orhan, O. Taştan, ALEVS: active learning by statistical leverage sampling, arXiv preprint arXiv:1507.04155(2015).

[6] J. Goetz, A. Tewari, P. Zimmerman, Active learning for non-parametric regression using purely random trees, Adv. Neural Inf. Process. Syst. 31 (2018).

[7] D. Wu, Pool-based sequential active learning for regression, IEEE Trans. Neural Netw. Learn. Syst. 30 (5) (2019) 1348–1359, doi:10.1109/TNNLS.2018.2868649.

[8] D. Wu, C.T. Lin, J. Huang, Active learning for regression using greedy sampling, Inf. Sci. 474 (2019) 90–105.

[9] D. Romero, M. Ma, G.B. Giannakis, Kernel-based reconstruction of graph signals, IEEE Trans. Signal Process. 65 (3) (2017) 764–778.

[10] P. Giménez-Febrer, A. Pagès-Zamora, G.B. Giannakis, Matrix completion and extrapolation via kernel regression, IEEE Trans. Signal Process. 67 (19) (2019) 5004–5017, doi:10.1109/TSP.2019.2932875.

[11] P. Giménez-Febrer, A. Pagès-Zamora, G.B. Giannakis, Generalization error bounds for kernel matrix completion and extrapolation, IEEE Signal Process. Lett. 27 (2020) 326–330, doi:10.1109/LSP.2020.2970306.

[12] P. Drineas, M.W. Mahoney, On the Nyström method for approximating a Gram matrix for improved kernel-based learning, J. Mach. Learn. Res. 6 (2005) 2153–2175.

[13] A. Alaoui, M.W. Mahoney, Fast randomized kernel ridge regression with statistical guarantees, in: Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 775–783. Montreal, Canada

[14] A. Tanaka, H. Imai, M. Miyakoshi, Kernel-induced sampling theorem, IEEE Trans. Signal Process. 58 (7) (2010) 3569–3577.

[15] C. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, in: Advances in Neural Information Processing Systems 13, MIT Press, 2001, pp. 682–688.
[16] C. Musco, C. Musco, Recursive sampling for the Nyström method, in: Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3833–3845.
[17] A. Deshpande, L. Rademacher, Efficient volume sampling for row/column subset selection, in: 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, IEEE, 2010, pp. 329–338.
[18] D.P. Woodruff, Sketching as a tool for numerical linear algebra, Theor. Comput. Sci. 10 (1–2) (2014) 1–157.
[19] F. Gama, A.G. Marques, G. Mateos, A. Ribeiro, Rethinking sketching as sampling: a graph signal processing approach, Signal Process. 169 (2020) 107404.
[20] S. Wang, A. Gittens, M.W. Mahoney, Sketched ridge regression: optimization perspective, statistical perspective, and model averaging, in: International Conference on Machine Learning, PMLR, 2017, pp. 3608–3616.
[21] R. Patel, T. Goldstein, E. Dyer, A. Mirhoseini, R. Baraniuk, Deterministic column sampling for low-rank matrix approximation: Nyström vs. incomplete Cholesky decomposition, in: Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 594–602.
[22] P. Drineas, R. Kannan, M.W. Mahoney, Fast Monte Carlo algorithms for matrices I: approximating matrix multiplication, SIAM J. Comput. 36 (1) (2006) 132–157.
[23] A. Rudi, R. Camoriano, L. Rosasco, Less is more: Nyström computational regularization, Advances in Neural Information Processing Systems 28 (2015).
[24] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, A. Zandieh, Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees, in: International conference on machine learning, PMLR, 2017, pp. 253–262.