Check for
updates

Nyström landmark sampling and regularized Christoffel functions

Michaël Fanuel[1] · **Joachim Schreurs[2]** · **Johan A. K. Suykens[2]**

Abstract
Selecting diverse and important items, called landmarks, from a large set is a problem of interest in machine learning. As a specific example, in order to deal with large training sets, kernel methods often rely on low rank matrix Nyström approximations based on the selection or sampling of landmarks. In this context, we propose a deterministic and a randomized adaptive algorithm for selecting landmark points within a training data set. These landmarks are related to the minima of a sequence of kernelized Christoffel functions. Beyond the known connection between Christoffel functions and *leverage scores*, a connection of our method with finite *determinantal point processes* (DPPs) is also explained. Namely, our construction promotes diversity among important landmark points in a way similar to DPPs. Also, we explain how our randomized adaptive algorithm can influence the accuracy of Kernel Ridge Regression.

Keywords Kernel · Nyström · Determinantal point process · Diversity

1 Introduction

Kernel methods have the advantage of being efficient, good performing machine learning methods which allow for a theoretical understanding. These methods make use of a kernel function that maps the data from input space to a kernel-induced feature space. Let $\{x_1, \ldots, x_n\}$ be a data set of points in $\mathbb{R}^d$ and let $k(x, y)$ be a strictly positive definite

✉ Michaël Fanuel
michael.fanuel@univ-lille.fr

Joachim Schreurs
joachim.schreurs@esat.kuleuven.be

Johan A. K. Suykens
johan.suykens@esat.kuleuven.be

1   Université de Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, Lille 59000, France

2   KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, Leuven 3001, Belgium

kernel. The associated kernel matrix is given by $K = [k(x_i, x_j)]_{i,j=1}^n$ and defines the similarity between data points. Just constructing the kernel matrix already requires $O(n^2)$ operations. Solving a linear system involving K, such as in kernel ridge regression, has a complexity $O(n^3)$. Therefore, improving the scalability of kernel methods has been an active research question. Two popular approaches for large-scale kernel methods are often considered: Random Fourier features (Rahimi & Recht, 2007) and the Nyström approximation (Williams & Seeger, 2001). The latter is considered more specifically in this paper. The accuracy of Nyström approximation depends on the selection of good landmark points and thus is a question of major interest. A second motivation, besides improving the accuracy of Nyström approximation, is selecting a good subset for regression tasks. Namely in Cortes et al. (2010) and Li et al. (2016b), it was shown that a better kernel approximation results in improving the performance of learning tasks. This phenomenon is especially important in stratified data sets (Chen et al., 2019; Oakden-Rayner et al., 2020; Valverde-Albacete & Peláez-Moreno, 2014). In these data sets, it is important to predict well in specific subpopulations or outlying points and not only in data dense regions. These outlying points can, e.g., correspond to serious diseases in a medical data set, being less common than mild diseases. Naturally, incorrectly classifying these outliers could lead to significant harm to patients. Unfortunately, the performance in these subpopulations is often overlooked. This is because aggregate performance measures such as MSE or sensitivity can be dominated by larger subsets, overshadowing the fact that there may be a small 'important' subset where performance is bad. These stratifications often occur in data sets with 'a long tail', i.e., the data distribution of each class is viewed as a mixture of distinct subpopulations (Feldman, 2020). Sampling algorithms need to select landmarks out each subpopulation to achieve a good generalization error. By sampling a *diverse* subset, there is a higher chance of every subpopulation being included in the sample. In this paper, the performance of learning tasks on stratified data is therefore measured on two parts: the bulk and tail of data. The bulk and the tail of data correspond to points with low and high outlyingness respectively, where outlyingness is measured by the ridge leverage scores which are described below.

1.1 Leverage score sampling

A commonly used approach for matrix approximation is the so-called Ridge Leverage Score (RLS) sampling. The RLSs intuitively correspond to the correlation between the singular vectors of a matrix and the canonical basis elements. In the context of large-scale kernel methods, recent efficient approaches indeed consider leverage score sampling (Bach, 2013; Drineas & Mahoney, 2005; Gittens & Mahoney, 2016) which results in various theoretical bounds on the quality of the Nyström approximation. Several refined methods also involve recursive sampling procedures (El Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2017a, 2018) yielding statistical guarantees. In view of certain applications where random sampling is less convenient, deterministic landmark selection based on leverage scores can also be used as for instance in Belabbas and Wolfe (2009), Papailiopoulos et al. (2014) and McCurdy (2018), whereas a greedy approach was also proposed in Farahat et al. (2011). While methods based on leverage scores often select 'important' landmarks, it is interesting for applications that require high sparsity to promote samples with 'diverse' landmarks since fewer landmarks correspond to fewer parameters for instance for Kernel Ridge Regression and LS-SVM (Suykens et al., 2002). This sparsity is crucial for certain (embedded) applications that often necessitate models with a

small number of parameters for having a *fast prediction* or due to memory limitations. In the same spirit, the selection of informative data points (Gauthier & Suykens, 2018; Guyon et al., 1996) is an interesting problem which also motivated entropy-based landmark selection methods (Girolami, 2002; Suykens et al., 2002).

1.2 Determinantal point processes

More recently, selecting diverse data points by using Determinantal Point Processes (DPPs) has been a topic of active research (Hough et al., 2006; Kulesza & Taskar, 2012a), finding applications in document or video summarization, image search taks or pose estimation (Gong et al., 2014; Kulesza & Taskar, 2010, 2011, 2012b). DPPs are ingenious probabilistic models of repulsion inspired by physics, and which are kernel-based. While the number of sampled points generated by a DPP is in general random, it is possible to define m-determinantal point processes (m-DPPs) which only generate samples of m points (Kulesza & Taskar, 2011). m-DPP-based landmark selection has shown to give superior performance for the Nyström approximation compared with existing approaches (Li et al., 2016a). However, the exact sampling of a DPP or a m-DPP from a set of n items takes $O(n^3)$ operations and therefore several improvements or approximations of the sampling algorithm have been studied in Derezinski et al. (2019), Li et al. (2016b) Li et al. (2016a) and Tremblay et al. (2019). In this paper, we show that our methods are capable of sampling equally or more diverse subsets as a DPP, which gives an interesting option to sample larger diverse subsets. Our paper explains how DPPs and leverage scores sampling methods can be connected in the context of Christoffel functions. These functions are known in approximation theory and defined in Appendix A. They have been 'kernelized' in the last years as a tool for machine learning (Askari et al., 2018; Lasserre & Pauwels, 2017; Pauwels et al., 2018) and are also interestingly connected to the problem of support estimation (Rudi et al., 2017b).

1.3 Roadmap

1. We start with an overview of RLS and DPP sampling for kernel methods. The connection between the two sampling methods is emphasized: both make use of the same matrix that we call here the projector kernel matrix. We also explain Christoffel functions, which are known to be connected to RLSs.
2. Motivated by this connection, we introduce conditioned Christoffel functions. These *conditioned* Christoffel functions are shown to be connected to *regularized* conditional probabilities for DPPs in Sect. 3, yielding a relation between Ridge Leverage Score sampling and DPP sampling.
3. Motivated by this analogy, two landmark sampling methods are analysed: Deterministic Adaptive Selection (DAS) and Randomized Adaptive Sampling (RAS) respectively in Sects. 4 and 5. Theoretical analyses of both methods are provided for the accuracy of the Nyström approximation. In particular, RAS is a simplified and weighted version of sequential sampling of a DPP (Launay et al., 2020; Poulson, 2020). Although slower compared to non-adaptive methods, RAS achieves a better kernel approximation and produces samples with a similar diversity compared to DPP sampling. The deterministic algorithm DAS is shown to be more suited to smaller scale data sets with a fast spectral decay.

4. Numerical results illustrating the advantages of our methods are given in Sect. 6. The proofs of our results are given in Appendix.

Notations Calligraphic letters ($\mathcal{C}, \mathcal{Y}, \ldots$) are used to denote sets with the exception of $[n] = \{1, \ldots, n\}$. Matrices ($A, K, \ldots$) are denoted by upper case letters, while random variables are written with boldface letters. Denote by $a \vee b$ the maximum of the real numbers a and b. The $n \times n$ identity matrix is $\mathbb{I}_n$ or simply $\mathbb{I}$ when no ambiguity is possible. The max norm and the operator norm of a matrix A are given respectively by $\|A\|_\infty = \max_{i,j} |A_{ij}|$ and $\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2$. The Moore-Penrose pseudo-inverse of a matrix A is denoted by A^+. Let A and B be symmetric matrices. We write $A \succeq B$ (resp. $A \succ B$) if $A - B$ is positive semidefinite (resp. positive definite). Let $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$. Then, if c is a real number and v a vector, $v \geq c$ denotes an entry-wise inequality. The i-th component of a vector v indicated by $[v]_i$, while $[A]_{ij}$ is the entry (i, j) of the matrix A. Given an $n \times n$ matrix A and $\mathcal{C} \subseteq [n]$, $A_{\mathcal{C}}$ denotes the submatrix obtained by selecting the columns of A indexed by $\mathcal{C}$, whereas the submatrix $A_{\mathcal{C}\mathcal{C}}$ is obtained by selecting both the rows and columns indexed by $\mathcal{C}$. The canonical basis of $\mathbb{R}^d$ is denoted by $\{e_\ell\}_{\ell=1}^d$.

Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with a strictly positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, written as $k \succ 0$. Typical examples of strictly positive definite kernels are given by the Gaussian RBF kernel $k(x, y) = e^{-\|x-y\|_2^2/(2\sigma^2)}$ or the Laplace kernel[1] $k(x, y) = e^{-\|x-y\|_2/\sigma}$. The parameter $\sigma > 0$ is often called the bandwidth of k. It is common to denote $k_x(\cdot) = k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathbb{R}^d$. In this paper, only the Gaussian kernel will be used. Finally, we write $f(x) \sim g(x)$ if $\lim_{x \to +\infty} f(x)/g(x) = 1$.

2 Kernel matrix approximation and determinantal point processes

Let K be an $n \times n$ positive semidefinite kernel matrix. In the context of this paper, landmarks are specific data points, within a data set, which are chosen to construct a low rank approximation of K. These points can be sampled independently by using some weights such as the ridge leverage scores or thanks to a point process which accounts for both importance and diversity of data points. These two approaches can be related in the framework of Christoffel functions.

2.1 Nyström approximation

In the context of Nyström approximations, the low rank kernel approximation relies on the selection of a subset of data points $\{x_i | i \in \mathcal{C}\}$ for $\mathcal{C} \subseteq [n]$. This is a special case of 'optimal experiment design' (Gao et al., 2019). For a small budget $|\mathcal{C}|$ of data points, it is essential to select points so that they are sufficiently diverse and representative of all data points in order to have a good kernel approximation. These points are called *landmarks* in order to emphasize that we aim to obtain both important and diverse data points. It is common to associate to a set of landmarks a sparse sampling matrix.

---

[1] By Bochner's theorem, a kernel matrix $K = [k(x_i, x_j)]_{i,j}$ of the Gaussian or Laplace kernel is non-singular almost surely, if the x_i's are sampled i.i.d. for instance from the Lebesgue measure.

Definition 1 (*Sampling matrix*) Let $\mathcal{C} \subseteq [n]$. A sampling matrix $S(\mathcal{C})$ is an $n \times |\mathcal{C}|$ matrix obtained by selecting and possibly scaling by a strictly positive factor the columns of the identity matrix which are indexed by $\mathcal{C}$. When the rescaling factors are all ones, the corresponding unweighted sampling matrix is simply denoted by C.

A simple connection between weighted and unweighted sampling matrices goes as follows. Let $S(\mathcal{C})$ be a sampling matrix. Then, there exists a vector of strictly positive weights $w \in (\mathbb{R}_{>0})^n$ such that $S(\mathcal{C}) = C \operatorname{Diag}(w)$, where C is an unweighted sampling matrix. An unweighted sampling matrix can be used to select submatrices as follows: let A be an $n \times n$ matrix and $\mathcal{C} \subseteq [n]$, then $A_{\mathcal{C}} = AC$ and $A_{\mathcal{CC}} = C^\top AC$. In practice, it is customary to use the unweighted sampling matrix to calculate the kernel approximation whereas the weighted sampling matrix is convenient to obtain theoretical results.

Definition 2 (*Nyström approximation*) Let K be an $n \times n$ positive semidefinite symmetric matrix and let $\mathcal{C} \subseteq [n]$. Given $\mu > 0$ and an $n \times |\mathcal{C}|$ sampling matrix $S(\mathcal{C})$, the *regularized* Nyström approximation of K is

$$L_{\mu,S}(K) = KS(S^\top KS + \mu \mathbb{I}_{|\mathcal{C}|})^{-1} S^\top K,$$

where $S = S(\mathcal{C})$. The *common* Nyström approximation is simply obtained in the limit $\mu \to 0$ as follows $L_{0,S}(K) = KS(S^\top KS)^+ S^\top K$.

The use of the letter L for the Nyström approximation refers to a Low rank approximation. The following remarks are in order.

- First, in this paper, the kernel is assumed to be strictly positive definite, implying that $K > 0$. In this case, the common Nyström approximation involves the inverse of the kernel submatrix in the above definition rather than the pseudo-inverse. Thus, the common Nyström approximation is then independent of the weights of the sampling matrix $S(\mathcal{C})$, i.e.,

$$L_{0,S(\mathcal{C})}(K) = L_{0,C}(K).$$

- Second, the Nyström approximation satisfies the structural inequality

$$0 \preceq K - L_{\mu,S}(K),$$

  for all sampling matrices S and for all $\mu \geq 0$. This can be shown thanks to Lemma 6 in Appendix.

- Third, if $S = C \operatorname{Diag}(w)$ with $w > 0$ and K is nonsingular, then we have

$$K - L_{0,C}(K) \preceq K - L_{\mu,S}(K),$$

  for all $\mu > 0$. The above inequality indicates that we can bound the error on the common Nyström approximation by upper bounding the error on the regularized Nyström approximation.

2.2 Projector kernel matrix

Ridge Leverage Scores and DPPs can be used to sample landmarks. The former allows guarantees with high probability on the Nyström error by using matrix concentration inequalities involving a sum of i.i.d. terms. The latter allows particularly simple error bounds on expectation as we also explain hereafter. Both these approaches and the algorithms proposed in this paper involve the use of another matrix constructed from K and that we call the *projector kernel matrix*, which is defined as follows

$$P_{n\gamma}(K) = K(K + n\gamma \mathbb{I})^{-1}, \tag{1}$$

where $\gamma > 0$ is a regularization parameter. This parameter is used to filter out small eigenvalues of K. When no ambiguity is possible, we will simply omit the dependence on $n\gamma$ and K to ease the reading. The terminology 'projector' can be understood as follows: if K is singular, then $P_{n\gamma}(K)$ is a mollification of the projector KK^+ onto the range of K.

We review the role of P for the approximation of K with RLS and DPP sampling in the next two subsections. These considerations motivate the algorithms proposed in this paper where P plays a central role.

2.3 RLS sampling

Ridge Leverage Scores are a measure of uniqueness of data points and are defined as follows

$$\ell_{n\gamma,K}(x_z) = [P_{n\gamma}(K)]_{zz}, \tag{2}$$

with $z \in [n]$. When no ambiguity is possible, the dependence on K is omitted in the above definition. RLS sampling is used to design approximation schemes in kernel methods as, e.g., in Musco and Musco (2017) and Rudi et al. (2017a). The trace of the projector kernel matrix is the so-called effective dimension (Zhang, 2005), i.e., $d_{\text{eff}}(\gamma) = \sum_{i \in [n]} \ell_{n\gamma}(x_i)$. For a given regularization $\gamma > 0$, this effective dimension is typically associated with the number of landmarks needed to have an 'accurate' low rank Nyström approximation of a kernel matrix as it is explained, e.g., in El Alaoui and Mahoney (2015).

To provide an intuition of the kind of guarantees obtained thanks to RLS sampling, we adapt an error bound from Calandriello et al. (2017a) for Nyström approximation holding with high probability, which was obtained originally in El Alaoui and Mahoney (2015) in a slightly different form. To do so, we firstly define the probability to sample $i \in [n]$ by

$$p_i = \ell_{n\gamma}(x_i)/d_{\text{eff}}(\gamma).$$

Then, RLS sampling goes as follows: one samples independently with replacement m landmarks indexed by $\mathcal{C} = \{i_1, \dots, i_m\} \subseteq [n]$ according to the discrete probability distribution given by p_i for $1 \le i \le n$. Then, the associated sampling matrix is $S = C \operatorname{Diag}(1/\sqrt{mp})$ with $p = (p_{i_1}, \dots, p_{i_m})^\top$.

Proposition 1 (El Alaoui and Mahoney (2015) *Thm. 2 and App. A, Lem. 1) Let $\gamma > 1/n$ be the regularization parameter. Let the failure probability be $0 < \delta < 1$ and fix a parameter $0 < t < 1$. If we sample m landmarks $\mathcal{C} = \{i_1, \dots, i_m\} \subseteq [n]$ with RLS sampling with*

$$m \geq \frac{2d_{\text{eff}}(\gamma) + 1/3}{t^2} \log\left(\frac{n}{\delta}\right)$$

and define the sampling matrix $S = C \operatorname{Diag}\left(1/\sqrt{mp}\right)$, then, with probability at least $1 - \delta$, we have

$$K - L_{n\gamma,S}(K) \preceq \frac{n\gamma}{1-t} P_{n\gamma}(K) \preceq \frac{n\gamma}{1-t} \mathbb{I}.$$

It is interesting to notice that the bound on Nyström approximation is improved if γ decreases, which also increases the effective dimension of the problem. To fix the ideas, it is common to take $t = 1/2$ in this type of results as it is done in El Alaoui and Mahoney (2015) and Musco and Musco (2017) for symmetry reasons within the proofs.

Remark 1 (*The role of column sampling and the projector kernel matrix*) The basic idea for proving Proposition 1 is to decompose the kernel matrix as $K = B^\top B$. Then, we have $P_{n\gamma}(K) = \Psi^\top \Psi$ with $\Psi = (BB^\top + n\gamma \mathbb{I})^{-1/2} B$ as it can be shown by the push-through identity given in Lemma 5 in Appendix. Next, we rely on an independent sampling of columns of Ψ which yields ΨS, where S is a weighted sampling matrix. Subsequently, a matrix concentration inequality, involving a sum of columns sampled from Ψ, is used to provide an upper bound on the largest eigenvalue of $\Psi\Psi^\top - \Psi SS^\top\Psi^\top$. Finally, the following key identity

$$K - L_{n\gamma,S}(K) = n\gamma \Psi^\top \left(\Psi SS^\top\Psi^\top + n\gamma(BB^\top + n\gamma\mathbb{I})^{-1}\right)^{-1}\Psi,$$

is used to connect this bound with Nyström approximation. A similar strategy is used within this paper to yield guarantees for RAS hereafter.

2.4 DPP sampling

Let us discuss the use of DPP sampling for Nyström approximation. We define below a specific class of DPPs, called L-ensembles, by following Hough et al. (2006) and Kulesza and Taskar (2012a).

Definition 3 (L-*ensemble*) Let L be an $n \times n$ symmetric positive-semidefinite matrix[2]. A finite Determinantal Point Process **Y** is an L-ensemble if the probability to sample the set $\mathcal{C} \subseteq [n]$ is given by

$$\Pr(\mathbf{Y} = \mathcal{C}) = \frac{\det(L_{\mathcal{C}\mathcal{C}})}{\det(\mathbb{I} + L)},$$

where $L_{\mathcal{C}\mathcal{C}}$ denotes the submatrix obtained by selecting the rows and columns indexed by $\mathcal{C}$. The corresponding process is denoted here by $\mathrm{DPP}_{\mathrm{L}}(L)$ where the subscript $_{\mathrm{L}}$ indicates it is an L-ensemble DPP, whereas the corresponding matrix L is specified as an argument.

In the above definition, the normalization factor is obtained by requiring that the probabilities sum up to unity. Importantly, for the specific family of matrices

---

[2] For a generalization to non-symmetric matrices, see e.g. Gartrell et al. (2019).

$$L = K/(n\gamma), \text{ such that } K \geq 0 \text{ is a kernel matrix and } \gamma > 0,$$

the marginal inclusion probabilities of this process $\Pr(\mathcal{C} \subseteq \mathbf{Y}) = \det(P_{\mathcal{C}\mathcal{C}})$ are associated with the following marginal (or correlation) kernel[3] $P = P_{n\gamma}(K)$, which is defined in (1). The expected sample size of $\text{DPP}_L(K/\alpha)$ for some $\alpha > 0$ is given by

$$\mathbb{E}_{\mathcal{C}\sim\text{DPP}_L(K/\alpha)}(|\mathcal{C}|) = \text{Tr}\left(P_\alpha(K)\right).$$

For more details, we refer to Kulesza and Taskar (2012a).

Let us briefly explain the repulsive property of L-ensemble sampling. It is well-known that the determinant of a positive definite matrix can be interpreted as a squared volume. A diverse sample spans a large volume and, therefore, it is expected to come with a large probability. For completeness, we give here a classical argument relating diversity and DPP sampling by looking at marginal probabilities. Namely, the probability that a random subset contains $\{i, j\}$ is given by

$$\Pr(\{i, j\} \subseteq \mathbf{Y}) = \det\begin{pmatrix} P_{ii} & P_{ij} \\ P_{ji} & P_{jj} \end{pmatrix} = \Pr(\{i\} \subseteq \mathbf{Y})\Pr(\{j\} \subseteq \mathbf{Y}) - P_{ij}^2.$$

So, if i and j are very dissimilar, i.e., if $|P_{ij}|$ or K_{ij} is small, the probability that i and j are sampled together is approximately the product of their marginal probabilities. Conversely, if i and j are very similar, the probability that they belong to the same sample is suppressed.

2.4.1 Nyström approximation error with L-ensemble sampling

We now explicitate the interest of landmark sampling with DPPs in the context of kernel methods. Sampling with L-ensemble DPPs was shown recently to achieve very good performance for Nyström approximation and comes with theoretical guarantees. In particular, the expected approximation error is given by the compact formula

$$\mathbb{E}_{\mathcal{C}\sim\text{DPP}_L(K/\alpha)}\left(K - L_{0,C}(K)\right) = \alpha P_\alpha(K) \preceq \alpha\mathbb{I}, \tag{3}$$

for all $\alpha > 0$ as shown independently in Fanuel et al. (2021) and Derezinski et al. (2020). In (3), C denotes the sampling matrix associated to the set $\mathcal{C}$; see Definition 1. The analogy with Proposition 1 is striking if we take $\alpha = n\gamma$. Again, a lower error bound on Nyström approximation error is simply $0 \preceq K - L_{0,C}(K)$ (this is a consequence of Lemma 6 in Supplementary Material F of Musco and Musco (2017)), which holds almost surely. The interpretation of the identity (3) is that the Nyström approximation error is controlled by the parameter $\alpha > 0$ which also directly influences the expected sample size $\text{Tr}\left(P_\alpha(K)\right)$. Indeed, a smaller value of α yields a larger expected sample size and gives therefore a more accurate Nyström approximation as explained in Fanuel et al. (2021). We want to emphasize that it is also common to define m-DPPs which are DPPs conditioned on a given subset size. While it is convenient to use m-DPPs in practice, their theoretical analysis is more involved; see Li et al. (2016b) and Schreurs et al. (2020).

---

[3] The marginal (or correlation) kernel of a DPP, usually denoted by K, is denoted here by P to prevent any confusion with the kernel matrix.

2.4.2 Sequential DPP sampling

Let $\mathcal{C}$ and $\tilde{\mathcal{C}}$ be disjoint subsets of $[n]$. The general conditional probability of a DPP both involves a condition of the type $\mathcal{C} \subseteq \mathbf{Y}$ and $\mathbf{Y} \cap \tilde{\mathcal{C}} = \emptyset$. The formula for the general conditional probability of a DPP is particularly useful for sequential sampling, which is also related to the sampling algorithms proposed in this paper.

It was noted recently in Poulson ([2020](#)) and Launay et al. ([2020](#)) that DPP $\mathbf{Y}$ sampling can be done in a sequential way. We consider here the case of an L-ensemble DPP. The sequential sampling of a DPP involves a sequence of n Bernoulli trials. Each Bernoulli outcome determines if one item is accepted or rejected. The algorithm loops over the set $[n]$ and constructs sequentially the disjoint sets $\mathcal{C}$ and $\tilde{\mathcal{C}}$ which are initially empty. At any step, the set $\mathcal{C}$ contains accepted items while $\tilde{\mathcal{C}}$ consists of items that were rejected. At step i, the element $i \in [n]$ is sampled if the outcome of a Bernoulli(p_i) random variable is successful. The success probability is defined as

$$p_i = \Pr(i \in \mathbf{Y} | \mathcal{C} \subseteq \mathbf{Y}, \mathbf{Y} \cap \tilde{\mathcal{C}} = \emptyset),$$

where $\mathbf{Y} \sim \mathrm{DPP_L}(L)$. Then, if i is sampled successfully, we define $\mathcal{C} \leftarrow \mathcal{C} \cup \{i\}$ and otherwise $\tilde{\mathcal{C}} \leftarrow \tilde{\mathcal{C}} \cup \{i\}$. The general conditional probability is given by

$$\Pr(z \in \mathbf{Y} | \mathcal{C} \subseteq \mathbf{Y}, \mathbf{Y} \cap \tilde{\mathcal{C}} = \emptyset) = [H^{\tilde{\mathcal{C}}} - L_{0,C}(H^{\tilde{\mathcal{C}}})]_{zz}, \tag{4}$$

where $\mathcal{C} \cap \tilde{\mathcal{C}} = \emptyset$ and the 'conditional marginal kernel' is

$$H^{\tilde{\mathcal{C}}} = P + P_{\tilde{C}}\left(\mathbb{I} - P_{\tilde{C}\tilde{C}}\right)^{-1} P_{\tilde{C}}^{\top}, \tag{5}$$

with $P = L(L + \mathbb{I})^{-1}$. The expression ([4](#)) can be found in a slightly different form in Launay et al. ([2020](#)) [Corollary 1]. From a computational perspective, P (or $\mathbb{I} - P$) may still have numerically very small eigenvalues so that the inverse appearing in the calculation of the Nyström approximation ([9](#)) or in ([5](#)) has to be regularized.

Exact sampling of a discrete DPP of a set of n items requires in general $O(n^3)$ operations. It is therefore interesting to analyse different sampling methods which can also yield diverse sampling, since exact DPP sampling is often prohibitive for large n. The role of DPP sampling for approximating kernel matrices motivates the next section where we extend the analogy between Christoffel functions and general conditional probabilities of L-ensemble DPPs.

3 Christoffel functions and determinantal point processes

3.1 Kernelized Christoffel functions

In the case of an empirical density, the value of the regularized Christoffel function at $x \in \mathbb{R}^d$ is obtained as follows:

$$C_\eta(x) = \inf_{f \in \mathcal{H}} \sum_{i=1}^{n} \frac{\eta_i}{n} f(x_i)^2 + \gamma \|f\|_{\mathcal{H}}^2 \text{ s.t. } f(x) = 1, \qquad \text{CF}$$

where $\eta_i > 0$ for $i \in [n]$ are pre-determined weights which can be used to give more or less emphasis to data points. A specific choice of weights is studied in Sect. [3.3](#). Notice that, in
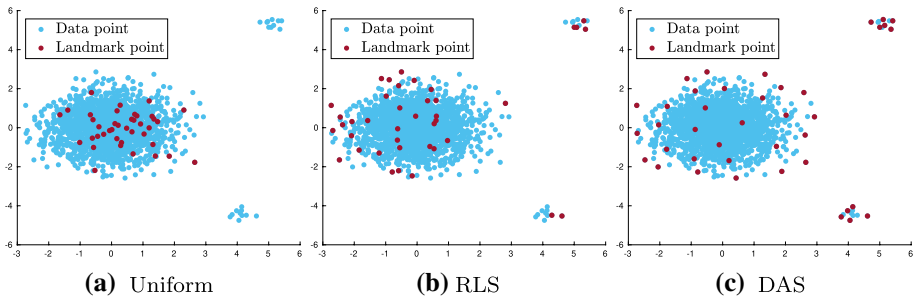
Fig. 1 Samplings of an artificial data set. Uniform sampling oversamples dense parts. RLS sampling overcomes this limitation while DAS (proposed in this paper) promotes more diversity.

the above definition, there is an implicit dependence on the parameter $\gamma > 0$, the empirical density and the RKHS itself. The functions defined by (CF) are related to the empirical density in the sense that $C_\eta(x)$ takes larger values where the data set is dense and smaller values in sparse regions. This is the intuitive reason why Christoffel functions can be used for outlier detection (Askari et al., 2018). As it is explained in Pauwels et al. (2018)[Eq. (2)], the Christoffel function takes the following value on the data set

$$C_\eta(x_i) = \left(e_i^\top K \left(K \operatorname{Diag}\left(\frac{\eta}{n}\right) K + \gamma K \right)^{-1} K e_i \right)^{-1} \quad \text{for all } i \in [n]. \tag{6}$$

The function defined by (CF) naturally provides importance scores of data points. The connection with RLSs (see (1) and (2)) when $\eta = 1$ can be seen thanks to the equivalent expression for $\eta > 0$

$$C_\eta(x_i) = \frac{\eta_i}{n} \left(e_i^\top K \left(K + n\gamma \operatorname{Diag}(\eta)^{-1} \right)^{-1} e_i \right)^{-1} \quad \text{for all } i \in [n], \tag{7}$$

which is readily obtained thanks to the matrix inversion lemma. For many data sets, the 'importance' of points is not uniform over the data set, which motivates the use of RLS sampling and our proposed methods as it is illustrated in Fig. 1. For completeness, we give a derivation of (6) in Appendix. Clearly, if $\eta > 0$, this Christoffel function can also be understood as an RLS associated to a reweighted kernel matrix $\bar{K}(\eta) = \operatorname{Diag}(\eta)^{1/2} K \operatorname{Diag}(\eta)^{1/2}$, i.e.,

$$C_\eta(x_i) = \frac{\eta_i}{n} \left([\bar{K}(\eta)(\bar{K}(\eta) + n\gamma \mathbb{I})^{-1}]_{ii} \right)^{-1} \tag{8}$$

for all $i \in [n]$.

In Sect. 3.2 below, we first define conditioned Christoffel functions and show their connection with a conditional probability for an L-ensemble DPP. This helps to provide some intuition and is a warm-up for the more general results of Sect. 3.3.

3.2 Conditioned Christoffel functions

We extend (CF) in order to provide importance scores of data points which are in the complement $\mathcal{C}^c$ of a set $\mathcal{C} \subseteq [n]$. This consists in excluding points of $\mathcal{C}$——and,

implicitly, to some extend also their neighbourhood—in order to look for other important points in $\mathcal{C}^c$. Namely, we define a conditioned Christoffel function as follows:

$$C_{\mathcal{C},\eta}(x_z) = \inf_{f \in \mathcal{H}} \sum_{i=1}^{n} \frac{\eta_i}{n} f(x_i)^2 + \gamma \|f\|_{\mathcal{H}}^2 \text{ s.t. } \begin{cases} f(x_z) = 1 \text{ and} \\ f(x_s) = 0 \ \forall s \in \mathcal{C} \end{cases}, \qquad \text{CondCF}$$

where we restrict here to x_z such that $z \in \mathcal{C}^c$. In order to draw a connection with leverage scores, we firstly take uniform weights $\eta_i = 1$ for all $i \in \mathcal{C}^c$. In this paper, we propose to solve several problems of the type (CondCF) for different sets $\mathcal{C} \subseteq [n]$. Then, we will select landmarks x_z which minimize $C_{\mathcal{C},\eta}(x_z)$ in the data set for a certain choice of η.

Now, we give a connection between Nyström approximation and conditioned Christoffel functions. Hereafter, Proposition 2 gives a closed expression for the conditioned Christoffel function in terms of the common Nyström approximation of the projector kernel matrix.

Proposition 2 *Let k be a strictly positive definite kernel and let $\mathcal{H}$ be the associated RKHS. For $\mathcal{C} \subset [n]$, we have*

$$C_{\mathcal{C},1}(x_z) = n^{-1}/[P - P_{\mathcal{C}} P_{\mathcal{C}\mathcal{C}}^{-1} P_{\mathcal{C}}^{\top}]_{zz}, \qquad (9)$$

for all $z \in \mathcal{C}^c$.

Clearly, in the absence of conditioning, i.e., when $\mathcal{C} = \emptyset$, one recovers the result of Pauwels et al. (2018) which has related the Christoffel function to the inverse leverage score. For $\mathcal{C} \neq \emptyset$, the second term at the denominator of (9) is clearly the common Nyström approximation (see Definition 2) of the projector kernel based on the subset of columns indexed by $\mathcal{C}$. Also, notice that $P - P_{\mathcal{C}} P_{\mathcal{C}\mathcal{C}}^{-1} P_{\mathcal{C}}^{\top}$ is the posterior predictive variance of a Gaussian process regression associated with the projector kernel P in the absence of regularization (Rasmussen & Williams, (2006)). The accuracy of the Nyström approximation based on $\mathcal{C}$ is given in terms of the max norm as follows:

$$\min_{z \in \mathcal{C}^c} C_{\mathcal{C},1}(x_z) = n^{-1}/\|P - L_{0,\mathcal{C}}(P)\|_{\infty},$$

where $L_{0,\mathcal{C}}$ is the common low rank Nyström approximation.

Now, we derive a connection between conditioned Christoffel functions and conditional probabilities for a specific DPP: an L-ensemble which is built thanks to a rescaled kernel matrix. To begin, we provide a result relating conditioned Christoffel functions and determinants in the simple setting of the last section. Below, Proposition 3, which can also be found in Belabbas and Wolfe (2009) in a slightly different setting, gives an equivalent expression for a conditioned Christoffel function.

Proposition 3 (Determinant formula) *Let $\mathcal{C} \subseteq [n]$ a non-empty set and $z \in \mathcal{C}^c$. Denote $\mathcal{C}_z = \mathcal{C} \cup \{z\}$. For $\eta = 1$, the solution of (CondCF) has the form:*

$$C_{\mathcal{C},1}(x_z) = n^{-1} \frac{\det P_{\mathcal{C}\mathcal{C}}}{\det P_{\mathcal{C}_z \mathcal{C}_z}}. \qquad (10)$$

Furthermore, let B be a matrix such that $P = B^{\top}B$. Then, we have the alternative expression $C_{\mathcal{C},1}(x_z) = n^{-1}/\|b_z - \pi_{V_{\mathcal{C}}} b_z\|_2^2$, where $b_z \in \mathbb{R}^{N \times 1}$ is the z-th column of B and $\pi_{V_{\mathcal{C}}}$ is the orthogonal projector on $V_{\mathcal{C}} = \text{span}\{b_s | s \in \mathcal{C}\}$.

Fig. 2 Christoffel function $C_{\eta(\mu,\nu)}(x_i)$ (blue line); see Proposition 4 for the precise definition. We take $\gamma = 0.01$ and the Gaussian kernel's bandwidth is $\sigma = 0.2/\sqrt{2}$. The data set is the following grid $x_i = -4 + 0.02i$ for all the integers $0 \le i \le 400$. The weights $\eta_i(\mu, \nu)$ for $\mu = 0.5$ and $\nu = 0.01$ are displayed in red whereas the corresponding sets are $\mathcal{C} = \{100\}$ and $\tilde{\mathcal{C}} = \{300\}$
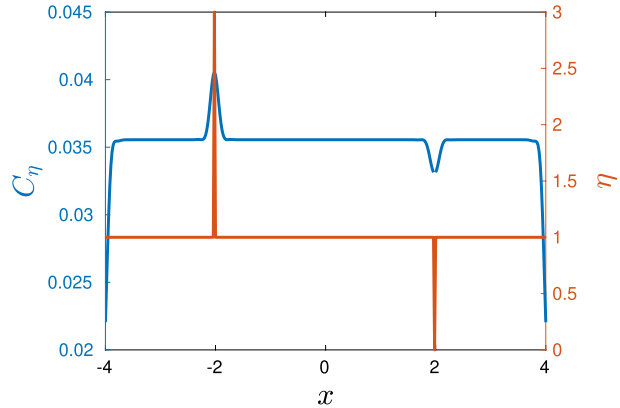
Interestingly, the determinants in (10) can be interpreted in terms of probabilities associated to an L-ensemble DPP with $L = K/(n\gamma)$. In other words, Proposition 3 states that the conditioned Christoffel function is inversely proportional to the conditional probability

$$\Pr(z \in \mathbf{Y}|\mathcal{C} \subseteq \mathbf{Y}) = \frac{\det P_{C_z C_z}}{\det P_{CC}}, \tag{11}$$

for $\mathbf{Y} \sim \mathrm{DPP}_L(K/(n\gamma))$. Hence, minimizing $C_{\mathcal{C},1}(x_z)$ over $z \in [n]$ produces an optimal $z^\star$ which is diverse with respect to $\mathcal{C}$. Inspired by the identity (11), we now provide a generalization in Sect. 3.3, which associates a common Christoffel function to a general conditional probability of a DPP by using a specific choice of weights η.

3.3 Regularized conditional probabilities of a DPP and Christoffel functions

Interestingly, the regularization of the expression of the general conditional probabilities for a DPP in (4) can be obtained by choosing specific weights η in the definition of the Christoffel function. To do so, the hard constraints in (CondCF) are changed to soft constraints in Proposition 4; see Fig. 2 for an illustration. In other words, the value of $f \in \mathcal{H}$ is severely penalized on x_i when $i \in \mathcal{C}$ in the optimization problem given in (CondCF). In the same way, f can also be weakly penalized on another disjoint subset $\tilde{\mathcal{C}}$.

Proposition 4 (Conditioning by weighting) *Let $\mu > 0$ and $0 < \nu < 1$. Let two disjoint subsets $\mathcal{C}, \tilde{\mathcal{C}} \subset [n]$ and define C and $\tilde{C}$ the matrices obtained by selecting the columns of the identity indexed by $\mathcal{C}$ and $\tilde{\mathcal{C}}$ respectively. Let $P = P_{n\gamma}(K)$ and define the following weights*:

$$\eta_i(\mu, \nu) = \begin{cases} 1 + \mu^{-1} & \text{if } i \in \mathcal{C}, \\ \nu & \text{if } i \in \tilde{\mathcal{C}}, \\ 1 & \text{otherwise}. \end{cases}$$

Then, the Christoffel function defined by (CF) associated to $\eta(\mu, \nu)$ is given by

$$C_{\eta(\mu,\nu)}(x_z) = n^{-1} / \left[H_\nu^{\tilde{\mathcal{C}}} - L_{\mu,\mathcal{C}}\left(H_\nu^{\tilde{\mathcal{C}}}\right) \right]_{zz},$$

where the regularized 'conditional marginal kernel' is

$$H_\nu^{\tilde{\mathcal{C}}} = P + (1-\nu)P_{\tilde{\mathcal{C}}}\left(\mathbb{I} - (1-\nu)P_{\tilde{\mathcal{C}}\tilde{\mathcal{C}}}\right)^{-1}P_{\tilde{\mathcal{C}}}^\top.$$

Hence, Proposition 4 shows that by choosing the weights η appropriately the corresponding Christoffel function gives a regularized formulation of conditional probabilities of a DPP if $\mu > 0$ and $0 < \nu < 1$ are small enough. Explicitly, we have

$$\lim_{\mu\to 0}\lim_{\nu\to 0}C_{\eta(\mu,\nu)}(x_z) = n^{-1}/\Pr(z\in\mathbf{Y}|\mathcal{C}\subseteq\mathbf{Y},\mathbf{Y}\cap\tilde{\mathcal{C}}=\emptyset),$$

for $\mathbf{Y} \sim \mathrm{DPP}_\mathrm{L}(K/(n\gamma))$ and $z \in [n]$, while the conditional probability on the RHS in given in (4). It is instructive to view the above Christoffel function as a Ridge Leverage Score of a reweighted matrix, bearing in mind the expression of (8). We emphasize that if μ and ν take small numerical values, the rows and columns of the reweighted matrix $\bar{K}$ indexed by a subset $\mathcal{C}$ have now very small values while the rows and columns indexed by $\tilde{\mathcal{C}}$ have large numerical values.

4 Deterministic adaptive landmark selection

It is often non trivial to obtain theoretical guarantees for deterministic landmark selection. We give here an approach for which some guarantees can be given. Inspired by the connection between DPPs and Christoffel functions, we propose to select deterministically landmarks by finding the minima of conditioned Christoffel functions defined such that the additional constraints enforce diversity. As it is described in Algorithm 1, we can successively minimize $C_{\mathcal{C}}(x_z)$ over a nested sequence of sets $\mathcal{C}_0 \subseteq \mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_m$ starting with $\mathcal{C}_0 = \emptyset$ by adding one landmark at each iteration. It is not excluded that the discrete maximization problem in Algorithm 1 can have several maxima. In practice, this is rarely the case. The ties can be broken uniformly at random or by using any deterministic rule. Hence, in the former case, the algorithm is not deterministic anymore.

Algorithm 1 DAS.

---

input: Matrix $K \succ 0$, sample size m and $\gamma > 0$.
initialization: $\mathcal{C}_0 = \emptyset$ and $l = 1$.
while: $l \leq m$ **do**
   $s_l \in \arg\max \mathrm{Diag}\left(P - P_{\mathcal{C}_l}P_{\mathcal{C}_l\mathcal{C}_l}^{-1}P_{\mathcal{C}_l}^\top\right).$
   $\mathcal{C}_l \leftarrow \mathcal{C}_{l-1} \cup \{s_l\}$ and $l \leftarrow l+1$.
end while
return $\mathcal{C}_m$.

In fact, Algorithm 1 is a greedy reduced basis method as it is defined in DeVore et al. (2013), which has been used recently for landmarking on manifolds in Gao et al. (2019) with a different kernel; see also Santin and Haasdonk (2017). Its convergence is studied in Proposition 5.

Proposition 5 (Convergence of DAS) *Let $\Lambda_1 \geq \cdots \geq \Lambda_n > 0$ be the eigenvalues of the projector kernel matrix P defined in (1). If $\mathcal{C}_m$ is sampled according to Algorithm 1, we have*:

$$\|P - P_{\mathcal{C}_m} P_{\mathcal{C}_m \mathcal{C}_m}^{-1} P_{\mathcal{C}_m}^{\top}\|_{\infty} \leq 2\|P\|_{\infty} \Lambda_{\lfloor m/2 \rfloor + 1}^{1/2}, \text{ for all } 2 \leq m < n.$$

Notice that $\|P\|_{\infty}$ is indeed the largest leverage score related to the maximal marginal degrees of freedom defined in Bach (2013). Furthermore, we remark that the eigenvalues of P are related to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$ of the kernel matrix K by the Tikhonov filtering $\Lambda_m = \frac{\lambda_m}{\lambda_m + n\gamma}$. In the simulations, we observe that DAS performs well when the spectral decay is fast. This seems to be often the case in practice as explained in Papailiopoulos et al. (2014) and McCurdy (2018). Hence, since DPP sampling is efficient for obtaining an accurate Nyström approximation, this deterministic adaptation can be expected to produce satisfactory results. However, it is more challenging to obtain theoretical guarantees for the accuracy of the Nyström approximation obtained from the selected subset. For a set of landmarks $\mathcal{C}$, Lemma 1 relates the error on the approximation of K to the error on the approximation of P.

Lemma 1 *Let $\mu > 0$ and $\tilde{\mu} = \mu/(\lambda_{max}(K) + n\gamma)$. Then, it holds that:*

$$0 \leq \frac{K - L_{\mu,S}(K)}{\lambda_{max}(K) + n\gamma} \leq P_{n\gamma}(K) - L_{\tilde{\mu},S}(P_{n\gamma}(K)).$$

This structural result connects the approximation of K to the approximation of P as a function of the regularization $n\gamma$. However, the upper bound on the approximation of K can be loose since $\lambda_{max}(K)$ can be large in general. It is also true that we might expect the Nyström approximation error to go to zero when the number of landmarks goes to n. Nevertheless, the rate of convergence to zero is usually unknown for deterministic landmark selection methods. For more instructive theoretical guarantees, we study a randomized version of Algorithm 1 in the next section.

Remark 2 (*Related approaches*) Several successful approaches for landmarking are based on Gaussian processes (Liang & Paisley, 2015) or use a Bayesian approach (Paisley et al., 2010) in the context of active learning. These methods are similar to DAS with the major difference that DAS uses the projector kernel $P = K(K + n\gamma\mathbb{I})^{-1}$, while the above mentioned methods rely on a kernel function $k(x, y)$ or its corresponding kernel matrix.

5 Randomized adaptive landmark sampling

RLS sampling and L-ensemble DPP sampling both allow for obtaining guarantees for kernel approximation as it is explained in Sect. 2. Randomized Adaptive Sampling (RAS) can be seen as an intermediate approach which gives a high probability guarantee for Nyström approximation error with a reduced computational cost compared with DPP sampling. We think RAS is best understood by considering the analogy with sequential DPP sampling and Christoffel functions.

As it is mentioned in Sect. 3.3, the sequential sampling of a DPP randomly picks items conditionally to the knowledge of previously accepted and rejected items, i.e., the disjoint sets $\mathcal{C}$ and $\tilde{\mathcal{C}}$ defined in Sect. 3.3. The computation of the matrix H in (5) involves solving a linear system to account for the past rejected samples. RAS uses a similar approach with the difference that the knowledge of the past rejected items is disregarded. Another difference with respect to DPP sequential sampling is that the sampling matrix used for

constructing the regularized Nyström approximation is weighted contrary to the analogous formula in (4). As explained in Algorithm 2, RAS constructs a sampling matrix by adaptively choosing to add (or not) a column to the sampling matrix obtained at the previous step.

Algorithm 2 RAS.

input: Matrix $K \succ 0$, oversampling $c > 0$ and regularizers $\gamma > 0$, $\epsilon \in (0, 1)$ and $t = 1/2$.
initialization: $S_0 = \emptyset$.
for i=1,..., n **do**
    calculate $s_i = \frac{1}{\epsilon} \left[P_{n\gamma}(K) - L_{\epsilon, S_{i-1}}(P_{n\gamma}(K)) \right]_{ii}$.
    calculate $p_i = \min\{1, c\tilde{l}_i\}$ with $\tilde{l}_i = \min\{1, (1+t)s_i\}$.
    draw $X_i \sim$ Bernoulli(p_i).
    if $X_i = 1$,
        add a column to the sampling matrix as follows $S_i \leftarrow (S_{i-1} | \frac{e_i}{\sqrt{p_i}})$.
    else
        $S_i \leftarrow S_{i-1}$ (no update).
    end if
end for
return S_n.

We now explain the sampling algorithm which goes sequentially over all $i \in [n]$. Initialize $\mathcal{S}_0 = \emptyset$. Let $i \in [n]$ and denote by S_{i-1} a sampling matrix associated to the set $\mathcal{S}_{i-1}$ at iteration $i - 1$. A Bernoulli random variable determines if i is added to the landmarks or not. If Bernoulli(p_i) is successful, we update the sampling matrix as follows $S_i = (S_{i-1} | \frac{e_i}{\sqrt{p_i}})$, otherwise $S_i = S_{i-1}$. The success probability p_i is defined to depend on the score

$$s_i = \frac{1}{\epsilon} \left[P_{n\gamma}(K) - L_{\epsilon, S_{i-1}}(P_{n\gamma}(K)) \right]_{ii},$$

by analogy with the conditional probability (11) and the Christoffel function value (9). The sampling probability is defined as

$$p_i = \min\{1, c\tilde{l}_i\} \quad \text{with} \quad \tilde{l}_i = \min\{1, (1+t)s_i\}, \tag{12}$$

with $t > 0$ and $c > 0$.

Remark 3 (*Oversampling parameter*) In this paper, the positive parameter c is considered as an oversampling factor and is chosen such that $c > 1$. This lower bound is less restrictive than the condition on c in Theorem 1. In our simulations, the oversampling parameter is always strictly larger than 1 so that the expression of the sampling probability simplifies to $p_i = \min\{1, c(1+t)s_i\}$.

Interpretation of the sampling probability. Here, $t > 0$ is a parameter that will take a fixed value and plays the same role as the parameter t in Proposition 1. Therefore t should not take a large value. Then, $\tilde{l}_i$ can be viewed as a leverage score. To allow for varying the number of sampled points, this leverage score is multiplied by an oversampling factor $c > 0$ as it is also done for recursive RLS sampling in Musco and Musco (2017). The

resulting score can be larger than 1 and therefore it is capped at 1. The idea is that the score s_i is an approximation of an RLS. Indeed, if we define the following factorization

$$B^\top B = P_{n\gamma}(K),\tag{13}$$

an alternative expression of the score is given by

$$s_i = [B^\top (BS_{i-1}S_{i-1}^\top B^\top + \epsilon \mathbb{I})^{-1}B]_{ii}$$

with $i \in [n]$, as a consequence of Lemma 6 in Appendix.

Then, the score s_i is an approximation of the ridge leverage score

$$[B^\top (BB^\top + \epsilon \mathbb{I})^{-1}B]_{ii} = [P_\epsilon(P_{n\gamma}(K))]_{ii},$$

for $i \in [n]$, with regularization parameter $\epsilon > 0$, as it can be shown thanks to the push-through identity[4]. In view of Sect. 2.3, the above ridge leverage score can be used for sampling a subset to compute a Nyström approximation of $P_{n\gamma}(K)$ thanks to Proposition 1. However, we are interested in the Nyström approximation of K. The following identity allows us to relate s_i to the approximation of K.

Lemma 2 *Let $K \succ 0$. Then, we have $P_\epsilon(P_{n\gamma}(K)) = \frac{1}{1+\epsilon}P_{\frac{\epsilon n\gamma}{1+\epsilon}}(K)$.*

The second regularization parameter ϵ is typically chosen as $\epsilon = 10^{-10}$ so that, in this circumstance, the adaptivity in Algorithm 2 promotes diversity among the landmarks in view of the discussion in Sect. 3.2.

Remark 4 (*Subset size*) Clearly, RAS outputs a subset of random size. This is also the case for L-ensemble DPP sampling, with the notable difference that the mean and variance of the subset size is known exactly in the latter case. For RAS, we empirically observe that the size variance is not large. The parameter $\gamma > 0$ influences the effective dimension of the problem and therefore helps to vary the amount of points sampled by RAS. A small γ yields a large effective dimension and conversely. Increasing c also increases the sample size which might also reduce diversity.

Theoretical guarantees. An error bound for Nyström approximation with RAS can be obtained by analogy with RLS sampling. Our main result in Theorem 1 deals with the adaptivity of RAS to provide an analogue of Proposition 1 which was designed for independent sampling.

We can now explain how RAS can be used to obtain a Nyström approximation of K by introducing an appropriate column sampling strategy. We define the factorization

$$P_\epsilon(P_{n\gamma}(K)) = \Psi^\top \Psi,$$

with $\Psi = (BB^\top + \epsilon \mathbb{I})^{-1/2}B$ and with B defined in (13). Therefore, thanks to Lemma 3 below, we know that if we can sample appropriately the columns of Ψ, denoted by $(\psi_i)_{i=1}^n$, then the error on the corresponding Nyström approximation will be bounded by a small constant.

---

[4] See Lemma 5 in Appendix.

Lemma 3 (Nyström approximation from column subset selection) *Let $\epsilon > 0$ and $0 < t < (1 + \epsilon)^{-1}$. Denote by Ψ a matrix such that $\Psi^\top \Psi = P_\epsilon(P_{n\gamma}(K))$. If there is a sampling matrix S such that*

$$\lambda_{max}(\Psi\Psi^\top - \Psi S S^\top \Psi^\top) \leq t,$$

then, it holds that

$$0 \leq K - L_{\frac{\epsilon n\gamma}{1+\epsilon}, S}(K) \leq \frac{\epsilon n\gamma}{1 - t(1 + \epsilon)} P_\epsilon(P_{n\gamma}(K)).$$

Some remarks are in order. Firstly, Lemma 3 has been adapted from El Alaoui and Mahoney (2015) in order to match our setting and its proof is given in Appendix for completeness. Secondly, the parameter t does not play a dominant role in the above error bound as long as it is not close to $(1 + \epsilon)^{-1}$. Indeed, the error bound can be simply improved by choosing a smaller value for the product $\epsilon n\gamma$. Hence, if we require $\epsilon < 1$, then we can choose $t = 1/2$ to clarify the result statement. Then, we have the simplified error bound

$$0 \leq K - L_{\frac{\epsilon n\gamma}{1+\epsilon}, S}(K) \leq \frac{2\epsilon n\gamma}{1 - \epsilon} \mathbb{I},$$

where we used that $P_\epsilon(P_{n\gamma}(K)) \preceq \mathbb{I}$ in Lemma 3.

In order to use this result, we want to find a good subset of columns of Ψ, i.e., a set of landmarks $\mathcal{S}$ and probabilities $(p_j)_{j \in \mathcal{S}}$ such that

$$\lambda_{max}\left(\sum_{i=1}^{n} \psi_i \psi_i^\top - \sum_{j \in \mathcal{S}} \psi_j \psi_j^\top / p_j \right) \leq t,$$

for $t > 0$, with high probability. This condition can be achieved for well-chosen probabilities $\{p_j : j \in \mathcal{S}\}$ by using a martingale concentration inequality which helps to deal with the adaptivity of RAS. This matrix martingale inequality plays the role of the matrix Bernstein inequality used in the proof of Proposition 1 in El Alaoui and Mahoney (2015).

Below, Theorem 1 indicates that, if the oversampling c is large enough, RAS produces a sampling matrix allowing for a 'good' Nyström approximation of K with high probability. For convenience, we recall the definition of the negative branch of the Lambert function $W_{-1} : [-e^{-1}, 0) \to \mathbb{R}$ which satisfies $\lim_{y \to 0_+} W_{-1}(-y)/\log(y) = 1$. Also, define the function $g(a) = -W_{-1}(-1/a)$ for all $a \geq e$, which has the asymptotic behaviour $g(a) \sim \log(a)$.

Theorem 1 (Main result) *Let $\epsilon \in (0, 1)$ and $\gamma > 0$. Let $\delta \in (0, 1)$. If the oversampling parameter satisfies:*

$$c \geq \frac{28}{3} g\left(\frac{700 d_{\text{eff}}(\frac{\epsilon n\gamma}{1+\epsilon})}{3(1 + \epsilon)\delta} \right) \vee \frac{1 + \sqrt{37}}{3},$$

then, with a probability at least $1 - \delta$, Algorithm 2 yields a weighted sampling matrix S such that

$$\|K - L_{\frac{\epsilon n\gamma}{1+\epsilon}, S}(K)\|_2 \leq \frac{2\epsilon n\gamma}{1 - \epsilon}.$$

Table 1 Small scale data sets and parameters

Data set	n	d	c/ϵ	σ
Housing	506	13	100	5
Stock	950	10	100	5
Abalone	4177	8	150	5
Bank 8FM	8192	8	150	10

Remark 5 (*Common Nyström approximation*) In view of the remarks of Sect. , the result of Theorem 1 directly implies a bound on the common Nyström approximation, that is $\|K - L_{0,S}(K)\|_2 \leq \frac{2\epsilon n\gamma}{1-\epsilon}$.

Discussion. Let us list a few comments about Theorem 1. The proof of Theorem 1, given in Appendix, uses the martingale-based techniques developed in Cohen et al. (2015) in the context of online sampling, together with a refinement by Minsker (2017) of the Freedman-type inequality for matrix martingales due to Tropp (2011). An advantage of the Freedman inequality obtained in Minsker (2017) is that it does not directly depend on the matrix dimensions but rather on a certain form of intrinsic dimension. The martingale-based techniques allow to deal with the adaptivity of our algorithm which precludes the use of statistical results using independent sampling. From the theoretical perspective, the adaptivity does not improve the type of guarantees obtained for independent sampling. The adaptivity of RAS rather makes the theoretical analysis more involved. However, our simulations indicates an empirical improvement. We emphasize that the dependence on the effective dimension in Theorem 1 does not appear in Cohen et al. (2015) although the proof technique is essentially the same. This extra dependence on $d_{\text{eff}}(\frac{\epsilon n\gamma}{1+\epsilon})$ helps to account for the kernel structure although the effective dimension can be however rather large in practice if $\epsilon \approx 10^{-10}$. A small enough ϵ indeed promotes diversity of sampled landmarks. The bound of Theorem 1 is pessimistic since it depends on the effective dimension which does not account for the subsample diversity. In its current state, the theory cannot yield improved guarantees over independent sampling. In practice, the chosen value for the oversampling factor is taken much smaller. A major difference with El Alaoui and Mahoney (2015) and Musco and Musco (2017) is that the landmarks $\mathcal{S}$ are *not* sampled independently in Algorithm 2. Algorithmically, RAS also corresponds to running Kernel Online Row Sampling (Calandriello et al., 2017b) with the difference that it is applied here on the projector kernel $P_{n\gamma}(K)$. As it is mentioned hereabove, a natural link can be made between RAS and sequential sampling of a DPP. RAS involves a regularized score $[P - L_{\epsilon,S_{i-1}}(P)]_{ii}$ at step i, where S_{i-1} is a (weighted) sampling matrix associated to a set $\mathcal{C}_{i-1} \subset [n]$. Indeed, this score depends on the points $\mathcal{C}_{i-1}$ which were chosen before step i, but not on the points $\tilde{\mathcal{C}}_{i-1}$ that were *not* chosen (corresponding to a failed Bernoulli). Sequential sampling of a DPP at step i also involves drawing a Bernoulli with probability $\Pr(i \in \mathbf{Y}|\mathcal{C}_{i-1} \subseteq \mathbf{Y}, \mathbf{Y} \cap \tilde{\mathcal{C}}_{i-1} = \emptyset)$ as given by a conditional kernel in (5). The first major difference with sequential sampling of a DPP is that we do *not* calculate the conditional kernel (5). The second difference is that the linear systems that are solved by RAS are *regularized* and Theorem 1 accounts for this regularization. Thirdly, RAS involves a *weighting* of the sampling matrix which does not appear in DPP sampling.

6 Numerical results

6.1 Exact algorithms

The performance of deterministic adaptive landmark sampling (DAS) and its randomized variant (RAS) is evaluated on the `Boston housing`, `Stock`, `Abalone` and `Bank 8FM` datasets which are described in Table 1.

These public data sets[5] have been used for benchmarking m-DPPs in Li et al. (2016b). The quality of the landmarks $\mathcal{C}$ is evaluated by calculating $\|K - \hat{K}\|_2/\|K\|_2$ with $\hat{K} = L_{\varepsilon,\mathcal{C}}(K)$ with $\varepsilon = 10^{-12}$ for numerical stability.

A Gaussian kernel is used with a fixed bandwidth $\sigma > 0$ after standardizing the data. We compare the following exact algorithms: Uniform sampling (Unif.), m-DPP,[6] exact ridge leverage score sampling (RLS) and DAS. The experiments have the following structure:

1. First, RAS is executed for an increasing $\gamma \in \{10^0, \dots, 10^{-4}\}$ for the `Boston housing` and `Stock` data sets, $\gamma \in \{10^0, \dots, 10^{-5}\}$ for the `Abalone` and `Bank 8FM` data sets and $\epsilon = 10^{-10}$. With each regularization parameter γ, RAS returns a different subset of m landmarks. The same subset size m is used for the comparisons with the other methods.
2. For each subset size m, the competing methods are executed for multiple hyperparameters $\gamma \in \{10^0, 10^{-1}, \dots, 10^{-6}\}$. The parameter γ here corresponds to the regularization parameter for the RLSs (see (2)), and the regularization parameter of DAS. The best performing γ is selected for each subset size m and sampling algorithm. In this manner, we each time compare the best possible sampling every algorithm can provide given the subset size m.

This procedure is repeated 10 times and the averaged results are visualized in Fig. 3. Notice that the error bars are small. DAS performs well on the `Boston housing` and `Stock` data set, which show a fast decay in the spectrum of K (see Fig. 10 in Appendix). This confirms the expectations from Proposition 5. If the decay of the eigenvalues is not fast enough, the randomized variant RAS has a superior performance which is similar to the performance of a m-DPP. For our implementation, RAS is faster compared with m-DPP, as it is illustrated in Fig. 13 in Appendix. The main cost of RAS is the calculation of the projector kernel matrix which requires $O(n^3)$ operations. The exact sampling of a m-DPP has a similar cost. As a measure of diversity the $\log\det(K_{\mathcal{CC}})$ is given in Fig. 11 in Appendix. The DAS method shows the largest determinant, RAS has a similar determinant as the m-DPP.

6.2 Approximate RAS

For medium-scale problems, we propose an approximate RAS method. First, the Gaussian kernel matrix K is approximated by using random Fourier features (Rahimi and Recht, 2007), that is, a generic method which is not data-adaptive. In practice, we calculate $F \in \mathbb{R}^{n \times n_F}$ with $n_F = 4000$ random Fourier features such that $FF^{\top}$ approximates K. Then, we obtain $\hat{P}_{n\gamma} = F(F^{\top}F + n\gamma\mathbb{I})^{-1}F^{\top}$, where the linear system is solved by using a Cholesky decomposition. This approximate projector matrix is used in the place of $P_{n\gamma}$

---

[5] https://www.cs.toronto.edu/%7edelve/data/datasets.html, https://www.openml.org/d/223.

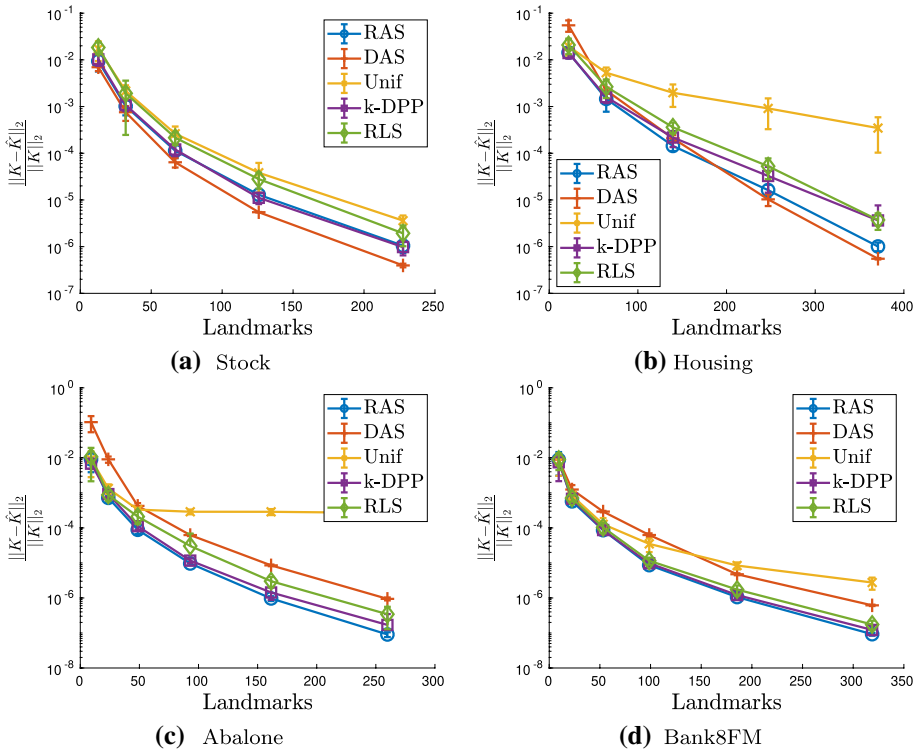[6] Matlab code: https://www.alexkulesza.com/.

Fig. 3 Relative operator norm of the Nyström approximation error as a function of the number of landmarks. The error is plotted on a logarithmic scale, averaged over 10 trials. The error bars are standard deviations.

in Algorithm 2. Next, we use the matrix inversion lemma in order to update the matrix inverse needed to calculate the sampling probability in Algorithm 2. The quality of the obtained landmarks is evaluated by calculating the averaged error $\|K_{\mathcal{A}\mathcal{A}} - \hat{K}_{\mathcal{A}\mathcal{A}}\|_F$ over 50 subsets $\mathcal{A} \subset [n]$ of 2000 points sampled uniformly at random. The Frobenius norm is preferred to the 2-norm to reduce the runtime. A comparison is done with the recursive leverage score sampling method[7] (RRLS) developed in Musco and Musco (2017) and BLESS which was proposed in Rudi et al. (2018). The parameters of BLESS are chosen in order to promote accuracy rather than speed (see Table 3 in Appendix). This algorithm was run several times until the number of sampled landmarks n_{BLESS} satifies $|\mathcal{C}| \leq n_{BLESS} \leq |\mathcal{C}| + 20$.

An overview of the data sets and parameters used in the experiments[8] is given in Table 2. Note that here we take a fixed σ and γ and not repeat the experiment for multiple subsets sizes $|\mathcal{C}|$.

The experiments are repeated 3 times, except for `Covertype`, for which the computation is done only once. The sampling with lowest average error over the 50 subsets is shown for each method. Hence, the error bars intuitively indicate the variance of the

---

Table 2 Medium scale data sets and parameters

Data set	n	d	c/ϵ	σ	γ
Super	21263	81	1	3	10^{-4}
CASP	45730	9	1	1	10^{-4}
Adult	48842	14	200	10	10^{-7}
MiniBooNE	130065	50	200	5,10	10^{-7}
cod-RNA	331152	8	200	4	10^{-6}
Covertype	581012	54	200	10	10^{-6}

(a) $|\mathcal{C}| = 2378, \sigma = 5$. **(b)** $|\mathcal{C}| = 924, \sigma = 10$.
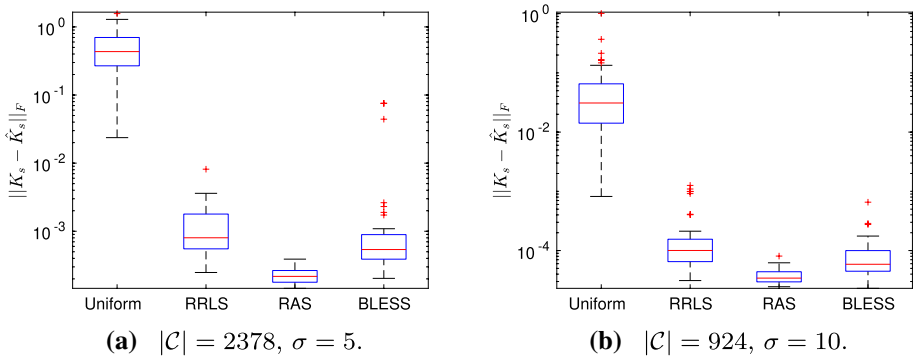
Fig. 4 Boxplot of the Nyström approximation error in Frobenius norm for MiniBooNE. The runtimes are: (a) $t_{RRLS} = 8$ s, $t_{RAS} = 710$ s, $t_{BLESS} = 18$ s and (b) $t_{RRLS} = 3$ s, $t_{RAS} = 167$ s, $t_{BLESS} = 5$ s.

(a) Adult, $|\mathcal{C}| = 1816$. **(b)** cod-RNA, $|\mathcal{C}| = 1280$. **(c)** Covertype, $|\mathcal{C}| = 3016$.
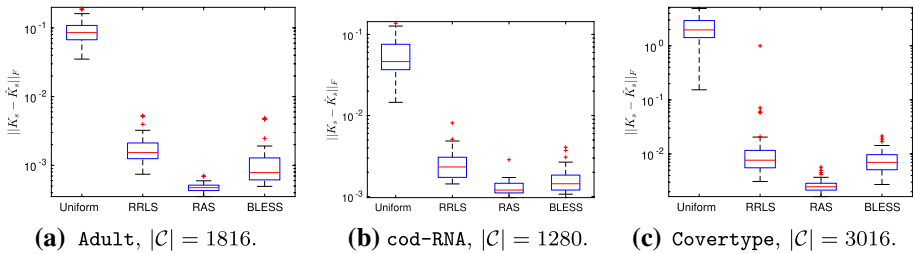
Fig. 5 Boxplot of the Nyström approximation error in Frobenius norm. The runtimes are: (a) $t_{RRLS} = 3$ s, $t_{RAS} = 197$ s, $t_{BLESS} = 4$ s, (b) $t_{RRLS} = 11$ s, $t_{RAS} = 555$ s, $t_{BLESS} = 11$ s and (c) $t_{RRLS} = 70$ s, $t_{RAS} = 3131$ s, $t_{BLESS} = 71$ s.

stochastic estimate of the Frobenius error. In Fig. 4, the sampling of landmarks with RAS for two different kernel parameters is compared to RRLS and BLESS on MiniBooNE. The boxplots show a reduction of the error, as well as a lower variance to the advantage of RAS. The increase in performance becomes more apparent when fewer landmarks are sampled. Additional experiments are given in Fig. 5 for Adult, cod-RNA and Covertype. The indicative runtimes for RRLS, RAS and BLESS are only given for completeness.
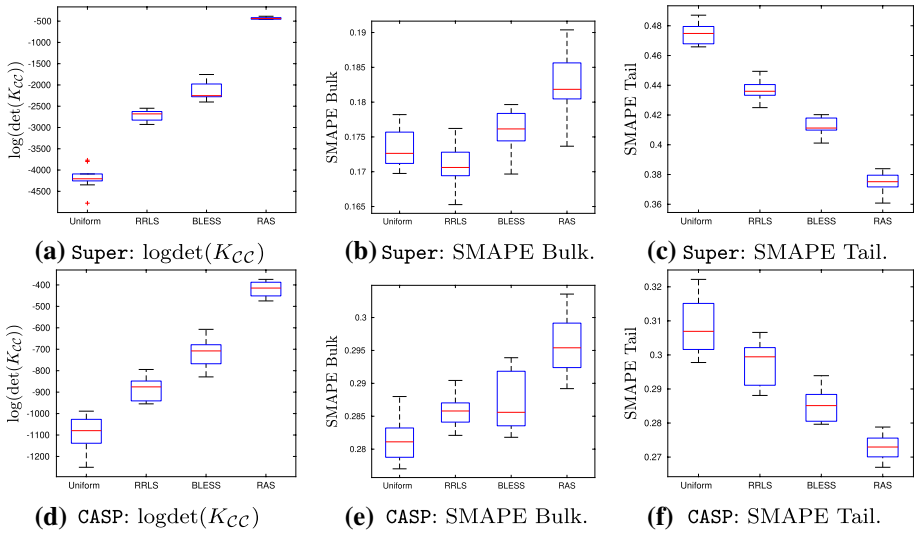
Fig. 6 Boxplot of the $\mathrm{logdet}(K_{CC})$, SMAPE in the bulk and tail. The runtimes are: (CASP) $t_{RRLS} = 0.4$ s, $t_{RAS} = 55$ s, $t_{BLESS} = 19$ s, (Super) $t_{RRLS} = 0.6$ s, $t_{RAS} = 68$ s, $t_{BLESS} = 14$ s. See also Fig. 7
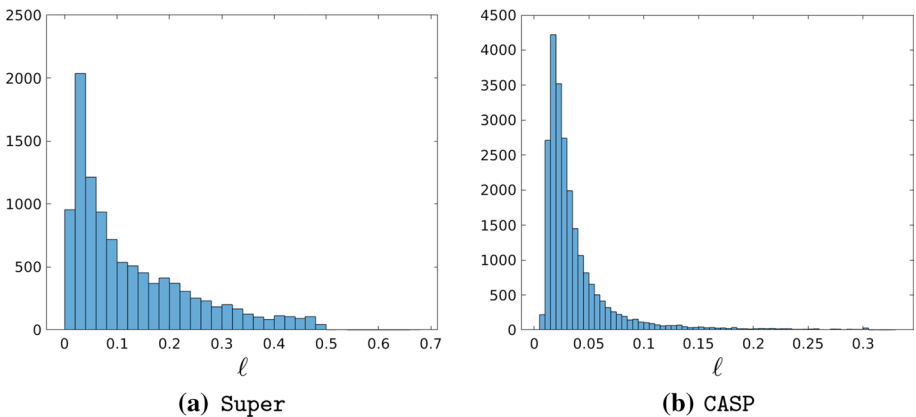

Fig. 7 The approximate ridge leverage scores accompanying Fig. 6

6.3 Performance for kernel ridge regression

To conclude, we verify the performance of the proposed method on a supervised learning task. Consider the input-output pairs $\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i \in [n]}$. The Kernel Ridge Regression (KRR) using the subsample $\mathcal{C}$ is determined by solving

$$f^\star = \arg\min_{f \in \mathcal{H}_\mathcal{C}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_\mathcal{H}^2, \text{ with } \lambda > 0,$$

where $\mathcal{H}_\mathcal{C} = \mathrm{span}\{k(x_i, \cdot) | i \in \mathcal{C}\}$. The regressor is $f^\star(\cdot) = \sum_{i \in \mathcal{C}} \alpha_i^\star k(x_i, \cdot)$ with

$$\alpha^{\star} = (K_{\mathcal{C}}^{\top} K_{\mathcal{C}} + n\lambda K_{\mathcal{C}\mathcal{C}})^{-1} K_{\mathcal{C}}^{\top} y. \tag{14}$$

The experiment is done on the following data sets: `Superconductivity` (`Super`) and `Physicochemical Properties of Protein Tertiary Structure` (`CASP`)[9], where we use a Gaussian Kernel after standardizing data. The symmetric mean absolute percentage error (SMAPE) between the outputs and the regressor

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - f^{\star}(x_i)|}{(|y_i| + |f^{\star}(x_i)|)/2},$$

is calculated for the methods: Uniform sampling, RRLS, BLESS and approximate RAS. First, approximate RAS is run with $n_F = 4000$ random Fourier features. Afterwards Uniform, RRLS and BLESS sample the same number of landmarks as approximate RAS. To evaluate the performance of different methods on the full space of data sets, the test set is stratified, i.e., the test set is divided into 'bulk' and 'tail' as follows. The bulk corresponds to test points where the RLSs are smaller than or equal to the 70% quantile, while the tail of data corresponds to test points where the ridge leverage score is larger than the 70% quantile. The RLSs are approximated using RRLS with 8000 data points. This way, we can evaluate if the performance of the different methods is also good for 'outlying' points. A similar approach was conducted in Fanuel et al. (2021). The data set is split in 50% training data and 50% test data, so to make sure the train and test set have similar RLS distributions. Information on the data sets and hyperparameters used for the experiments is given in Table 2. The regularization parameter $\lambda \in \{10^{-4}, 10^{-6}, 10^{-8}, 10^{-12}\}$ of Kernel Ridge Regression is determined by using 10-fold-cross-validation. The results in Fig. 6 show that approximate RAS samples a more diverse subset. As a result, the method performs much better in the tail of the data, while yielding a slightly worse performance in the bulk. The performance gain in the tail of the data is more apparent for the `Super`, where the RLS distribution has a longer tail. The data set has more 'outliers', increasing the importance of using a diverse sampling algorithm. The RLS distributions of the data sets are visualized in Appendix.

7 Conclusion

Motivated by the connection between leverage scores, DPPs and kernelized Christoffel functions, we propose two sampling algorithms: DAS and RAS, with a theoretical analysis. RAS allows to sample diverse and important landmarks with a good overall performance. An additional approximation is proposed so that RAS can be applied to large data sets in the case of a Gaussian kernel. Experiments for kernel ridge regression show that the proposed method performs much better in the tail of the data, while maintaining a comparable performance for the bulk data.

---

[9] https://archive.ics.uci.edu/ml/datasets.php.

Appendix A: Christoffel function and orthogonal polynomials

For completeness, we simply recall here the definition of the Christoffel function (Pauwels et al., 2018). Let $\ell \in \mathbb{N}$ and $p(x)$ be an integrable real function on $\mathbb{R}^d$. Then, the Christoffel function reads:

$$z \mapsto \min_{P \in \mathbb{R}_\ell(X)} \int P^2(x)p(x)\mathrm{d}x \text{ s.t. } P(z) = 1,$$

where $\mathbb{R}_\ell(X)$ is the set of d variate polynomials of degree at most ℓ.

In comparison, the optimization problem (CF) involves a minimization over an RKHS and includes an additional regularization term. Let $f \in \mathcal{H}$ and $\alpha \in \mathbb{R}^n$. Then, denote the sampling operator $\mathfrak{S} : \mathcal{H} \to \mathbb{R}^n$, such that $\mathfrak{S}f = \frac{1}{\sqrt{n}}(f(x_1) \ldots f(x_n))^\top$ and its adjoint $\mathfrak{S}^* : \mathbb{R}^n \to \mathcal{H}$ given by $\mathfrak{S}^*\alpha = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i k_{x_i}$. In order to rephrase the optimization problem in the RKHS, we introduce the covariance operator $\mathfrak{S}^*\mathfrak{S} = \frac{1}{n} \sum_{i=1}^n k_{x_i} \otimes k_{x_i}$ and the kernel matrix $\mathfrak{S}\mathfrak{S}^* = \frac{1}{n}K$ which share the same non-zero eigenvalues. Then, the problem (CondCF) reads:

$$C_{\mathcal{C}}(x_z) = \inf_{f \in \mathcal{H}}\langle f, (\mathfrak{S}^*\mathfrak{S} + \gamma \mathbb{I})f \rangle_{\mathcal{H}} \text{ s.t. } \langle k_{x_z}, f \rangle_{\mathcal{H}} = 1 \text{ and } f \in (\mathrm{span}\{k_{x_s} | s \in \mathcal{C}\})^\perp. \quad (15)$$

In other words, the formulation (15) corresponds to the definition (CF) where the RKHS $\mathcal{H}$ is replaced by a specific subspace of $\mathcal{H}$.

Appendix B: useful results and deferred proofs

In Sect. B.1, some technical lemmata are gathered, while the next subsections provide the proofs of the results of this paper.

B.1 Useful results

We firstly give in Lemma 4 a result which is used to compute the expression of conditioned Christoffel functions in Proposition 2.

Lemma 4 *Let H be a Hilbert space. Let $\{v_s \in H | s \in \mathcal{C} \subseteq [n]\}$ a collection of vectors and denote $V = \mathrm{span}\{v_s \in H | s \in \mathcal{C} \subseteq [n]\}$. Define $\pi_{V^\perp}$ the linear projector onto the orthogonal of V. Let $u \in H$ such that $\pi_{V^\perp}u \neq 0$. Then, the optimal value of*

$$m^\star = \min_{x \in H}\langle x, x \rangle \text{ s.t. } \begin{cases} \langle x, u \rangle = 1 \\ \langle x, v_s \rangle = 0, \text{ for all } s \in \mathcal{C} \end{cases} \quad (16)$$

is given by $m^\star = 1/\|\pi_{V^\perp}u\|^2$.

Proof Let $x = x_\perp + x_\parallel$ where $x_\perp = \pi_{V^\perp}x$ and $x_\parallel = \pi_V x$. Then, the objective satisfies

$$\langle x, x \rangle = \langle x_\perp, x_\perp \rangle + \langle x_\parallel, x_\parallel \rangle \geq \langle x_\perp, x_\perp \rangle.$$

This indicates that we can simply optimize over $x \in V^\perp$. The problem is then

$$m^\star = \min_{x \in V^\perp} \langle x, x \rangle \text{ s.t. } \langle x, \pi_{V^\perp} u \rangle = 1.$$

By using a Lagrange multiplier $\lambda \in \mathbb{R}$, we find $2x + \lambda \pi_{V^\perp} u = 0$. Next, we use the constraint $\langle x, \pi_{V^\perp} u \rangle = 1$. Thus, we find that the multiplier has to satisfy $\lambda = -2/\|\pi_{V^\perp} u\|^2$. After a substitution, this yields finally $x^\star = \pi_{V^\perp} u / \|\pi_{V^\perp} u\|^2$. $\qquad\square$

Next, we recall in Lemma 5 an instrumental identity which is often called 'push-through identity'.

Lemma 5 (Push-through identity) *Let X and Y be $n \times m$ matrices such that $X^\top Y + \mathbb{I}_m$ and $YX^\top + \mathbb{I}_n$ are invertible. Then, we have*

$$(X^\top Y + \mathbb{I}_m)^{-1} X^\top = X^\top (YX^\top + \mathbb{I}_n)^{-1}.$$

The following identity is a slightly generalized version of a result given in Musco and Musco (2017), which originally considers only the diagonal elements of the matrix $K - L_{\mu,S}(K)$ below.

Lemma 6 (Lemma 6 in Appendix F of Musco and Musco (2017)) *Let $\mu > 0$ and S be a sampling matrix. Let B such that $K = B^\top B$. We have*

$$K - L_{\mu,S}(K) = \mu B^\top (BSS^\top B^\top + \mu \mathbb{I})^{-1} B.$$

Proof This is a consequence of the push-through identity in Lemma 5. $\qquad\square$

In what follows, we list two results in Lemma 7 and Lemma 8 which are used in the proof of Proposition 4.

Lemma 7 *Let B be an $n \times n$ matrix. Let S be an $n \times k$ matrix and $\alpha \in \mathbb{R}$, such that $\mathbb{I} + \alpha B^\top SS^\top B$ and $\mathbb{I} + \alpha S^\top BB^\top S$ are non-singular. Then, we have*

$$B(\mathbb{I} + \alpha B^\top SS^\top B)^{-1} B^\top = BB^\top - \alpha BB^\top S(\mathbb{I} + \alpha S^\top BB^\top S)^{-1} S^\top BB^\top.$$

Proof We consider the expression on the RHS. It holds that

$$\begin{aligned}
& BB^\top - \alpha BB^\top S(\mathbb{I} + \alpha S^\top BB^\top S)^{-1} S^\top BB^\top \\
&= B(\mathbb{I} - \alpha B^\top S(\mathbb{I} + \alpha S^\top BB^\top S)^{-1} S^\top B)B^\top \\
&= B(\mathbb{I} - \alpha B^\top SS^\top B(\mathbb{I} + \alpha B^\top SS^\top B)^{-1})B^\top = B(\mathbb{I} + \alpha B^\top SS^\top B)^{-1} B^\top,
\end{aligned}$$

where the next to last equality is due to the push-through identity in Lemma 5 with $X = \alpha S^\top B$ and $Y = S^\top B$. This yields the desired result. $\qquad\square$

Lemma 8 *Let $K \succeq 0$ and let M be $n \times n$ symmetric matrix such that $K + M \succ 0$. Define $P_M = K^{1/2}(K + M)^{-1} K^{1/2}$. Let S be an $n \times k$ matrix and $\alpha \in \mathbb{R}$ be such that $K + M + \alpha K^{1/2} SS^\top K^{1/2}$ and $\mathbb{I} + \alpha S^\top P_M S$ are non-singular. Then, we have*

$$K^{1/2}\big(K + M + \alpha K^{1/2} SS^\top K^{1/2}\big)^{-1} K^{1/2} = P_M - \alpha P_M S\big(\mathbb{I} + \alpha S^\top P_M S\big)^{-1} S^\top P_M.$$

Proof The results follows from Lemma 7 by choosing $B = K^{1/2}(K + M)^{-1/2}$ such that $BB^\top = P_M$. $\qquad\square$

B.2 Proof of Proposition 2

Proof We start by setting up some notations. Let $f \in \mathcal{H}$ and $\alpha \in \mathbb{R}^n$. Then, define the sampling operator $\mathfrak{S} : \mathcal{H} \to \mathbb{R}^n$, such that $\mathfrak{S}f = \frac{1}{\sqrt{n}}(f(x_1) \dots f(x_n))^\top$. Its adjoint writes $\mathfrak{S}^* : \mathbb{R}^n \to \mathcal{H}$ and is given by $\mathfrak{S}^* \alpha = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i k_{x_i}$.

Firstly, since $z \in X$, the objective (CondCF) and the constraints depend only on the discrete set $X = \{x_1, \dots, x_n\}$. Then, we can decompose $f = f_X + f_X^\perp$ where $\langle f_X, f_X^\perp \rangle_{\mathcal{H}} = 0$ and $f_X \in \text{span}\{k_{x_i} | i \in [n]\}$. Since, by construction, $f_X^\perp(x_i) = 0$ for all $i \in [n]$, the objective satisfies:

$$\frac{1}{n} \sum_{i=1}^n f_X(x_i)^2 + \gamma \|f\|_{\mathcal{H}}^2 \geq \frac{1}{n} \sum_{i=1}^n f_X(x_i)^2 + \gamma \|f_X\|_{\mathcal{H}}^2.$$

Hence, we can rather minimize on $f \in \text{span}\{k_{x_i} | i \in [n]\}$. This means that we can find $\alpha \in \mathbb{R}^n$ such that $f^\star = \sum_{i=1}^n \alpha_i k_{x_i} = \sqrt{n}\mathfrak{S}^*\alpha$. The objective (CondCF) reads

$$\begin{aligned}
\langle f^\star, (\mathfrak{S}^*\mathfrak{S} + \gamma \mathbb{I})f^\star \rangle_{\mathcal{H}} &= \sqrt{n}\alpha^\top \mathfrak{S}(\mathfrak{S}^*\mathfrak{S} + \gamma \mathbb{I})\mathfrak{S}^*\alpha \sqrt{n} \\
&= n\alpha^\top(\mathfrak{S}\mathfrak{S}^* + \gamma \mathbb{I})\mathfrak{S}\mathfrak{S}^*\alpha \\
&= n\alpha^\top(K/n + \gamma \mathbb{I})(K/n)\alpha \\
&= n^{-1}\alpha^\top(K + n\gamma \mathbb{I})K\alpha.
\end{aligned}$$

For simplicity, denote by $K_{x_i} = Ke_i$ a column of the matrix K, where e_i is the i-th element of the canonical basis of $\mathbb{R}^n$. The problem becomes then:

$$\min_{\alpha \in \mathbb{R}^n} \alpha^\top M \alpha \text{ such that } \alpha^\top K_{x_z} = 1 \text{ and } \alpha^\top K_{x_s} = 0 \text{ for all } s \in \mathcal{C},$$

where $M = n^{-1}K(K + n\gamma \mathbb{I})$. Recall that, since we assumed $k > 0$, then K is invertible. By doing the change of variables $\alpha_M = M^{1/2}\alpha$, we obtain:

$$\min_{\alpha_M \in \mathbb{R}^n} \alpha_M^\top \alpha_M \text{ such that } \alpha_M^\top K_{x_z}^{(M)} = 1 \text{ and } \alpha_M^\top K_{x_s}^{(M)} = 0 \text{ for all } s \in \mathcal{C},$$

with $K_{x_z}^{(M)} = M^{-1/2}K_{x_z}$ and $K_{x_s}^{(M)} = M^{-1/2}K_{x_s}$. Let $V = \text{span}\{K_{x_s}^{(M)} | s \in \mathcal{C}\}$ and $K^{(M)} \in \mathbb{R}^{n \times |\mathcal{C}|}$ be the matrix whose columns are the vectors $K_{x_s}^{(M)}$. Notice that the elements of $\{K_{x_s}^{(M)} | s \in \mathcal{C}\}$ are linearly independent since $K_{x_s}^{(M)} = M^{-1/2}Ke_s$. Then, we use Lemma 4, and find $C_{\mathcal{C}}^{-1}(x_z) = \|\pi_{V^\perp} K_{x_z}^{(M)}\|^2$. By using $\pi_V = K^{(M)}G^{-1}K^{(M)\top}$ with the Gram matrix $G = K^{(M)\top}K^{(M)}$, we find:

$$\|\pi_{V^\perp} K_{x_z}^{(M)}\|^2 = K_{x_z}^{(M)\top}\left(\mathbb{I} - K^{(M)}G^{-1}K^{(M)\top}\right)K_{x_z}^{(M)}. \tag{17}$$

By substituting $M = n^{-1}K(K + n\gamma \mathbb{I})$, we find that $G_{ss'} = ne_s^\top K(K + n\gamma \mathbb{I})^{-1}e_{s'}$ for all $s, s' \in \mathcal{C}$. Hence, we find:

$$\|\pi_{V^\perp} K_{x_z}^{(M)}\|^2 = n\{e_z^\top K(K + n\gamma \mathbb{I})^{-1}e_z - e_z^\top K(K + n\gamma \mathbb{I})^{-1}G^{-1}K(K + n\gamma \mathbb{I})^{-1}e_z\},$$

which yields the desired result. $\square$

B.3 Details of derivations of Section 3.1

By using a similar approach and the same notations as in the previous subsection, we find

$$C_\eta(x_z) = \min_{\alpha \in \mathbb{R}^n} \alpha^\top \left(\frac{1}{n} K \operatorname{Diag}(\eta) K + \gamma K \right) \alpha \text{ such that } \alpha^\top K_{x_z} = 1,$$

for some $z \in [n]$. By using Lemma 4 we find the expression given in (6)

$$C_\eta(x_z) = \left(K_{x_z}^\top \left(\frac{1}{n} K \operatorname{Diag}(\eta) K + \gamma K \right)^{-1} K_{x_z} \right)^{-1} = \frac{n^{-1}}{[K(K \operatorname{Diag}(\eta) K + n\gamma K)^{-1} K]_{zz}}.$$

The second expression for Christoffel functions given in (7) is merely obtained by using the matrix inversion lemma as follows

$$K(K \operatorname{Diag}(\eta) K + n\gamma K)^{-1} K = K \left(\frac{K^{-1}}{n\gamma} - \left(\operatorname{Diag}(\eta)^{-1} + \frac{K}{n\gamma} \right)^{-1} \right) K$$

$$= \frac{1}{n\gamma} \left(K - K(K + n\gamma \operatorname{Diag}(\eta))^{-1} \right)$$

$$= K \left(K + n\gamma \operatorname{Diag}(\eta)^{-1} \right)^{-1} \operatorname{Diag}(\eta)^{-1}.$$

By taking the (z, z)-th entry of the matrix above gives directly (7).

B.4 Proof of Proposition 3

Proof The objective $C_C(x_z)^{-1}$ can be simply obtained by calculating the following determinant

$$n^{-1} C_C(x_z)^{-1} = \det \left(\begin{pmatrix} P_{zz} & P_{z,C} \\ P_{z,C}^\top & P_{C,C} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0^\top & P_{C,C} \end{pmatrix}^{-1} \right)$$

$$= \det \begin{pmatrix} P_{zz} & P_{z,C} P_{C,C}^{-1} \\ P_{z,C}^\top & \mathbb{I} \end{pmatrix} = P_{zz} - P_{z,C} P_{C,C}^{-1} P_{z,C}^\top,$$

where we use the well-known formula for the determinant of block matrices at the last equality. This yields the desired result. □

B.5 Proof of Proposition 4

Proof By using Lemma 4 or the results of Pauwels et al. (2018), we find the expression

$$n^{-1}(e_i^\top \underbrace{K(K \operatorname{Diag}(\eta) K + n\gamma K)^{-1} K}_{T} e_i)^{-1} = \inf_{f(x_z)=1} \frac{1}{n} \sum_{i=1}^n \eta_i f(x_i)^2 + \gamma \|f\|_{\mathcal{H}}^2, \quad (18)$$

where the weights are defined as $\eta_i = 1 + \mu^{-1}$ for all $i \in C$ and $\eta_i = \nu$ if $i \in \tilde{C}$ and $\eta_i = 1$ otherwise. Recall that $\mu > 0$ and $0 < \nu < 1$ so that $\operatorname{Diag}(\eta)$ is the positive diagonal matrix

$$\mathrm{Diag}\,(\eta) = \mathbb{I} + \mu^{-1}CC^{\top} + (\nu - 1)\tilde{C}\tilde{C}^{\top},$$

where C (resp. $\tilde{C}$) is a sampling matrix obtained by sampling the columns of the identity matrix belonging to $\mathcal{C}$ (resp. $\tilde{\mathcal{C}}$). Notice that $CC^{\top}$ (resp. $\tilde{C}\tilde{C}^{\top}$) is the diagonal matrices with entries equal to 1 for elements indexed by $\mathcal{C}$ (resp. $\tilde{\mathcal{C}}$) and equal to zero otherwise. First, we establish an equivalent expression for T at the LHS of (18), as follows

$$
\begin{aligned}
T &= K(K\,\mathrm{Diag}\,(\eta)K + n\gamma K)^{-1}K \\
&= K^{1/2}\big(K^{1/2}\big(\mathbb{I} + \mu^{-1}CC^{\top} + (\nu - 1)\tilde{C}\tilde{C}^{\top}\big)K^{1/2} + n\gamma\mathbb{I}\big)^{-1}K^{1/2}.
\end{aligned}
$$

Then, the idea is to define the symmetric matrix $M = (\nu - 1)K^{1/2}\tilde{C}\tilde{C}^{\top}K^{1/2} + n\gamma\mathbb{I}$, which is such that

$$K + M = K^{1/2}(\mathbb{I} - \tilde{C}\tilde{C}^{\top})K^{1/2} + \nu K^{1/2}\tilde{C}\tilde{C}^{\top}K^{1/2} + n\gamma\mathbb{I} > 0, \tag{19}$$

since $\mathbb{I} - \tilde{C}\tilde{C}^{\top} \geq 0$. Then, define the following notation (also used in Lemma 8):

$$P_M = K^{1/2}(K + M)^{-1}K^{1/2} = K^{1/2}\big(K + n\gamma\mathbb{I} - (1 - \nu)K^{1/2}\tilde{C}\tilde{C}^{\top}K^{1/2}\big)^{-1}K^{1/2},$$

which satisfies $P_M > 0$ since $K + M > 0$ and $K > 0$. Next, by using Lemma 8 with $\alpha = \mu^{-1}$ and the matrix M defined above (19), we find the following equivalent expression for T defined in (18), i.e.,

$$
\begin{aligned}
T &= K^{1/2}\left(K + \underbrace{(\nu - 1)K^{1/2}\tilde{C}\tilde{C}^{\top}K^{1/2} + n\gamma\mathbb{I}}_{M} + \mu^{-1}K^{1/2}CC^{\top}K^{1/2}\right)^{-1}K^{1/2} \\
&= P_M - \mu^{-1}P_M C\big(\mathbb{I} + \mu^{-1}C^{\top}P_M C\big)^{-1}C^{\top}P_M.
\end{aligned}
$$

Now, in order to match the notation of Proposition 4, we define $H = P_M$. This gives

$$T = H - \mu^{-1}HC\big(\mathbb{I} + \mu^{-1}C^{\top}HC\big)^{-1}C^{\top}H = H - L_{\mu,C}(H). \tag{20}$$

Next, to find an equivalent expression for

$$H = K^{1/2}\big(K + n\gamma\mathbb{I} + (\nu - 1)K^{1/2}\tilde{C}\tilde{C}^{\top}K^{1/2}\big)^{-1}K^{1/2},$$

we use once more Lemma 8, but this time we take $\alpha = \nu - 1$ and $M = n\gamma\mathbb{I}$. Notice that α does not need to be positive to be able to apply Lemma 8. An equivalent expression for H is

$$H = P + (1 - \nu)P\tilde{C}\big(\mathbb{I} - (1 - \nu)\tilde{C}^{\top}P\tilde{C}\big)^{-1}\tilde{C}^{\top}P, \tag{21}$$

with $P = K(K + n\gamma\mathbb{I})^{-1}$. Remark that the assumptions of Lemma 8 are satisfied: on the one hand, the matrix inverse in (21) exits since $0 \leq \tilde{C}^{\top}P\tilde{C} \leq \mathbb{I}$ and $0 < \nu < 1$; on the other hand, the matrix $K + n\gamma\mathbb{I} + (\nu - 1)K^{1/2}\tilde{C}\tilde{C}^{\top}K^{1/2}$ is non-singular in the light of (19). By combining (20) and (21), we have the desired result. $\qquad\square$

Appendix C: Proof of the results of Section 4

C.1 Proof of Proposition 5

Proof Let us explain how Algorithm 1 can be rephrased by considering first Proposition 3. First, we recall the spectral factorization of the projector kernel matrix $P = B^\top B$. For simplicity, denote the Hilbert space[10] $H = \mathbb{R}^n$. We also introduce the notation $\mathcal{F} = \{b_1, \dots, b_n\}$ for the set of columns of $B \in \mathbb{R}^{n \times n}$ which is considered as a compact subset of H. The vector subspace generated by the columns indexed by the subset $\mathcal{C}_m$ is denoted by $V_m = \mathrm{span}\{b_s | s \in \mathcal{C}_m\} \subset H$. With these notations and thanks to Proposition 3, the *square root* of the objective maximized by Algorithm 1 can be written $\sigma_m = n^{-1/2} C_{\mathcal{C}_m}^{-1/2}(x_{s_{m+1}}) = \|b_{s_{m+1}} - \pi_{V_m} b_{s_{m+1}}\|_2$. Hence, by using the definition of the Christoffel function, we easily see that $\sigma_{m+1} \le \sigma_m$. Then, the algorithm can be viewed as a greedy reduced subspace method, namely:

$$\sigma_m = \max_{b \in \mathcal{F}} \mathrm{dist}(b, V_m),$$

where the distance between a column $b \in \mathcal{F}$ and the subspace V_m is given by:

$$\mathrm{dist}(b, V_m) = \min_{y \in V_m} \|b - y\|_2. \tag{22}$$

As explained in DeVore et al. (2013); Santin and Haasdonk (2017); Gao et al. (2019); Binev et al. (2011), the convergence of this type of algorithm can be studied as a function of the Kolmogorov width:

$$d_m = \min_{Y_m} \max_{b \in \mathcal{F}} \mathrm{dist}(b, Y_m), \tag{23}$$

where the minimization is over all m-dimensional vector subspace Y_m of H. Intuitively, d_m is the best approximation error calculated over all possible subspace of dimension m with $m < n$. Then, Theorem 3.2 together with Corollary 3.3 in DeVore et al. (2013) gives the upper bound:

$$\sigma_{2m} \le \sqrt{2\|P\|_\infty d_m}, \quad \text{for all } m \ge 1, \tag{24}$$

which is a close analogue of eq. (4.9) in Gao et al. (2019). Hence, by upper bounding the Kolmogorov width, we provide a way to calculate a convergence rate of the greedy algorithm. To do so, we first recall that $\mathcal{F} = \{b_1, \dots b_n\}$ where the i-th column of B is denoted by $b_i = Be_i$. Therefore, we have:

$$\max_{b \in \mathcal{F}} \min_{y \in Y_m} \|b - y\|_2 = \max_{i \in [n]} \min_{y \in Y_m} \|Be_i - y\|_2 \le \max_{\|x\|_2 \le 1} \min_{y \in Y_m} \|Bx - y\|_2, \tag{25}$$

where the inequality is obtained by noticing that e_i is such that $\|e_i\|_2 = 1$ and by observing that maximizing an objective over a larger set yields a larger objective value. Then, in the light of this remark and according to Pinkus (1979), we define a modified width as follows:

$$\tilde{d}_m(B) = \min_{Y_m} \max_{\|x\|_2 \le 1} \min_{y \in Y_m} \|Bx - y\|_2,$$

---

[10] Here, we do not endow H with an RKHS structure.

which naturally satisfies $d_m \leq \tilde{d}_m(B)$ as it can be seen by taking the minimum of both sides of (25) over the m-dimensional vector spaces Y_m and by recalling the Kolmogorov width definition in (23). Then, $\tilde{d}_m(B)$ can be usefully upper bounded thanks to Theorem 2 that we adapt from Pinkus (1979).

Theorem 2 (Thm 1.1 in Pinkus (1979)) *Let B be a real $p \times q$ matrix. Let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ denote the p eigenvalues of $BB^\top$ and let $v_1, \ldots, v_p$ denote a corresponding set of orthonormal eigenvectors. Then, $\tilde{d}_m(B) = \lambda_{m+1}^{1/2}$ if $m < p$ and $\tilde{d}_m(B) = 0$ if $m \geq p$. Furthermore if $m < p$, then an optimal subspace for $\tilde{d}_m$ is $\mathrm{span}\{v_1, \ldots, v_m\}$.*

Our convergence result given in Corollary 5 simply follows from (24), the inequality $d_m \leq \tilde{d}_m$ and Theorem 2 by noting that the non-zero eigenvalues of $BB^\top$ are the non-zero eigenvalues of $B^\top B$. Notice that B is obtained by the spectral factorization of $P = B^\top B$ and therefore B is full rank. $\qquad\square$

C.2 Proof of Lemma 1

Proof To prove this result, we rely on Lemma 6 and on the definition of $P = K(K + n\gamma\mathbb{I})^{-1}$. Firstly, by following Lemma 6, we have

$$K - L_{\mu,S}(K) = \mu K^{1/2}(K^{1/2}SS^\top K^{1/2} + \mu\mathbb{I})^{-1}K^{1/2}, \qquad (26)$$

with $\mu > 0$. Then, we can obtain a similar expression for the projector kernel

$$\begin{aligned}
P - L_{\mu,S}(P) &= \mu P^{1/2}(P^{1/2}SS^\top P^{1/2} + \mu\mathbb{I})^{-1}P^{1/2} \\
&= \mu K^{1/2}\big(K^{1/2}SS^\top K^{1/2} + \mu(K + n\gamma\mathbb{I})\big)^{-1}K^{1/2},
\end{aligned}$$

where we substituted the definition $P^{1/2} = K^{1/2}(K + n\gamma\mathbb{I})^{-1/2}$ in the last equality. Next, we use $K \preceq \lambda_{max}(K)\mathbb{I}$, and the property that $A^{-1} \succeq B^{-1}$ if $B \succeq A$ with A and $B \succ 0$. We obtain:

$$\big(K^{1/2}SS^\top K^{1/2} + \mu(K + n\gamma\mathbb{I})\big)^{-1} \succeq \big(K^{1/2}SS^\top K^{1/2} + \mu(\lambda_{max}(K) + n\gamma)\mathbb{I}\big)^{-1},$$

and the result follows from the following inequality

$$\begin{aligned}
P - L_{\mu,S}(P) &\succeq \mu K^{1/2}\big(K^{1/2}SS^\top K^{1/2} + \mu(\lambda_{max}(K) + n\gamma)\mathbb{I}\big)^{-1}K^{1/2} \\
&= (\lambda_{max}(K) + n\gamma)^{-1}\tilde{\mu}K^{1/2}\big(K^{1/2}SS^\top K^{1/2} + \tilde{\mu}\mathbb{I}\big)^{-1}K^{1/2} \\
&= (\lambda_{max}(K) + n\gamma)^{-1}\big(K - L_{\tilde{\mu},S}(K)\big),
\end{aligned}$$

where we used again (26) in the last equality and defined $\tilde{\mu} = \mu(\lambda_{max}(K) + n\gamma)$. $\qquad\square$

Appendix D: Proof of the results of Section 5

D.1 Proof of Lemma 3

The statement of Lemma 3 has been adapted from El Alaoui and Mahoney (2015) to the setting of this paper. Its proof, given for completeness, follows the same lines as in Lemma 1 in the appendix of El Alaoui and Mahoney (2015).

Proof For simplicity let $\theta = \epsilon n\gamma/(1 + \epsilon)$. By assumption, we have

$$\Psi\Psi^\top - \Psi SS^\top \Psi^\top \preceq t\mathbb{I},$$

with $P_\epsilon(P_{n\gamma}(K)) = \Psi^\top \Psi$. By using the composition rule given in Lemma 2, we find the equivalent expression

$$\Psi^\top \Psi = P_\epsilon(P_{n\gamma}(K)) = \frac{1}{1 + \epsilon}P_\theta(K) \text{ with } \theta = \frac{\epsilon n\gamma}{1 + \epsilon}.$$

Therefore, we can choose the factorization given by $\Psi = (1 + \epsilon)^{-1/2}(K + \theta\mathbb{I})^{-1/2}K^{1/2}$. By using our starting hypothesis and by denoting for simplicity $P_\theta = P_\theta(K)$, we have

$$\frac{1}{1 + \epsilon}\left(P_\theta - P_\theta^{1/2}SS^\top P_\theta^{1/2}\right) \preceq t\mathbb{I}.$$

After a substitution of $P_\theta(K) = K(K + \theta\mathbb{I})^{-1}$, this gives the following inequality

$$(K + \theta\mathbb{I})^{-1/2}\left(K - K^{1/2}SS^\top K^{1/2}\right)(K + \theta\mathbb{I})^{-1/2} \preceq t(1 + \epsilon)\mathbb{I}.$$

Equivalently, we have the following inequality

$$K - K^{1/2}SS^\top K^{1/2} \preceq t(1 + \epsilon)(K + \theta\mathbb{I}),$$

which gives

$$K^{1/2}SS^\top K^{1/2} \succeq K - t(1 + \epsilon)(K + \theta\mathbb{I}). \tag{27}$$

Then, if $1 - t(1 + \epsilon) > 0$, it holds that

$$
\begin{aligned}
K - L_{\theta,S}(K) &= \theta K^{1/2}\left(K^{1/2}SS^\top K^{1/2} + \theta\mathbb{I}\right)^{-1}K^{1/2} && \text{(by Lemma 6)} \\
&\preceq \theta K^{1/2}\left(K + \theta\mathbb{I} - t(1 + \epsilon)(K + \theta\mathbb{I})\right)^{-1}K^{1/2} && \text{(by using (27))} \\
&= \frac{\theta}{1 - t(1 + \epsilon)}P_\theta(K),
\end{aligned}
$$

where we used $P_\theta(K) = K(K + \theta\mathbb{I})^{-1}$ at the last equality. Now, by using Lemma 2, we find

$$K - L_{\frac{\epsilon n\gamma}{1+\epsilon},S}(K) \preceq \frac{\epsilon n\gamma/(1 + \epsilon)}{1 - t(1 + \epsilon)}(1 + \epsilon)P_\epsilon(P_{n\gamma}(K)),$$

which is the stated result.　　　　　　　　　　　　　　　　　　　　　　　　　　　□

D.2 Proof of Theorem 1

The proof structure follows closely Cohen et al. (2015) with suitable adaptations. In particular and in contrast with Cohen et al. (2015), a main ingredient of the proof of Theorem 1 is the following improved Freedman inequality.

Theorem 3 (Thm 3.2 in Minsker (2017)) *Let $X_1, \ldots, X_n$ be martingale differences valued in the $n \times n$ symmetric matrices and let $U > 0$ be a deterministic real number such that $\|X_i\|_2 \leq U$ almost surely. Denote the variance $W_n = \sum_{i=1}^{n} \mathbb{E}_{i-1}(X_i^2)$. Then, for any $t \geq \frac{1}{6}(U + \sqrt{U^2 + 36\sigma^2})$, we have:*

$$
\Pr\left(\| \sum_{i=1}^{n} X_i \|_2 > t, \lambda_{max}(W_n) \leq \sigma^2 \right) \leq 50 \, \mathrm{Tr}\left(\min\{ \frac{t}{U} \frac{\mathbb{E}(W_n)}{\sigma^2}, 1\} \right) \exp\left(\frac{-t^2/2}{\sigma^2 + tU/3} \right).
$$
(28)

Notice that the function $\min\{M, 1\}$, defined for all symmetric $M \succeq 0$, acts on the eigenvalues of M. In Theorem 3, σ^2 denotes an upper bound on the predictable quadratic variation and not the kernel bandwidth parameter.

We now give the proof of Theorem 1.

Proof Let $t > 0$ and $0 < \epsilon < 1$. We consider here the sampling procedure given by RAS; see Algorithm 2. We recall that we defined the matrix B such that $B^\top B = P_{n\gamma}(K)$. Then, this proof considers the problem of sampling columns of the matrix Ψ which is defined such that

$$
\Psi = (BB^\top + \epsilon\mathbb{I})^{-1/2} B.
$$

Proof strategy The key idea of this proof is to define a matrix martingale $Y_0, \ldots, Y_n$ involving a sum of rank 1 matrices built from the columns of Ψ so that, at step n, we have

$$
Y_n = \Psi S_n S_n^\top \Psi^\top - \Psi\Psi^\top,
$$

where S_n is the sampling matrix at step n. If RAS samples enough landmarks, i.e., if S_n samples enough columns of Ψ, then Y_n will be small with high probability. In other words, we want to show that for a large enough $c > 0$, we have

$$
\|Y_n\|_2 \leq t,
$$

and therefore $\lambda_{max}(Y_n) \leq t$ with a large probability. Then, by using Lemma 3 and by choosing $t = 1/2$, we can obtain an error bound on a regularized Nyström approximation

$$
0 \preceq K - L_{\frac{\epsilon n\gamma}{1+\epsilon}, S}(K) \preceq \frac{2\epsilon n\gamma}{1 - \epsilon}\mathbb{I},
$$

which is the result that we want to prove.

The proof goes as follows. The matrix martingale is firstly defined. In order to use the martingale concentration inequality, we upper bound the norm of the difference sequence which boils down to find a lower bound on the sampling probability. Finally, we use the concentration inequality find a lower bound on c so that $\|Y_n\|_2 \leq t$ with a large probability.

Defining a matrix martingale Let S_i be the sampling matrix at step i in Algorithm 2 and p_i be the sampling probability. Define now a matrix martingale $Y_0, Y_1, \ldots, Y_n \in \mathbb{R}^{n \times n}$ such that $Y_0 = 0$ and with difference sequence $X_1, \ldots, X_n$ given by:

$$X_i = \begin{cases} (\frac{1}{p_i} - 1)\psi_i\psi_i^\top & \text{if column } b_i \text{ is sampled,} \\ -\psi_i\psi_i^\top & \text{otherwise,} \end{cases} \tag{29}$$

where p_i is given in (12) and where ψ_i is the i-th column of Ψ. So, we have $X_i = Y_i - Y_{i-1}$ for $1 \le i \le n$. Notice also that if $p_i = 1$, then $X_i = 0$. It remains to show that we indeed defined a martingale. To do so, we define $\mathbb{E}_{i-1}$ as the expectation conditioned on the previous steps from 0 to $i - 1$ included, more precisely

$$\mathbb{E}_{i-1}[Z] = \mathbb{E}[Z | Y_0, X_1, \ldots, X_{i-1}],$$

for some random Z. We simply have $\mathbb{E}_{i-1}(X_i) = (p_i(\frac{1}{p_i} - 1) - (1 - p_i))\psi_i\psi_i^\top = 0$ by a direct calculation.

Next, denote by $B_{[i-1]}$ the rectangular matrix obtained by selecting the $i - 1$ first columns of B, so that $B_{[n]} = B$. So, the martingale value at step $i - 1$ reads

$$Y_{i-1} = (BB^\top + \epsilon\mathbb{I})^{-1/2}\left(BS_{i-1}S_{i-1}^\top B^\top - B_{[i-1]}B_{[i-1]}^\top\right)(BB^\top + \epsilon\mathbb{I})^{-1/2}.$$

We now describe a modification of the martingale allowing for using Theorem 3 by identifying the constants U and σ^2 in the tail bound.

We first observe that, if $\|Y_{i-1}\|_2 < t$ at step $i - 1$, then by construction $\|Y_j\|_2 < t$ for all previous steps $j < i - 1$. If, at some point $\|Y_{i-1}\|_2 \ge t$, then we set $X_j = 0$ for all subsequent steps $j \ge i$. By abusing notations, we denote this stopped process by the same symbol $(Y_i)_i$; for more details about stopped martingales, we refer to Tropp (2019)[Section 7]. The basic idea is that, if the norm of the stopped martingale at step $i - 1$ is strictly smaller than t, the same is true for the original martingale defined in (29). The stopped martingale facilitates the analysis, since as soon as it has stopped the difference sequence satisfies $X_i = 0$. Again, if the martingale has not yet stopped $\|Y_{i-1}\|_2 < t$ and X_i is given by (29), and we have

$$(BB^\top + \epsilon\mathbb{I})^{-1/2}\left(BS_{i-1}S_{i-1}^\top B^\top - B_{[i-1]}B_{[i-1]}^\top\right)(BB^\top + \epsilon\mathbb{I})^{-1/2} \prec t\mathbb{I},$$

which gives the following inequality

$$BS_{i-1}S_{i-1}^\top B^\top \prec B_{[i-1]}B_{[i-1]}^\top + t(BB^\top + \epsilon\mathbb{I}). \tag{30}$$

Notice that, by definition, B contains the columns of $B_{[i-1]}$, and therefore, the following inequality holds

$$B_{[i-1]}B_{[i-1]}^\top \preceq BB^\top. \tag{31}$$

Lower bounding the sampling probability At this point, we want to show that the norm of the difference sequence X_i is bounded in order to use the matrix martingale concentration result which is given below in Theorem 3. We have the elementary upper bound on the difference sequence

$$\|X_i\|_2 \leq \max\{1, 1/p_i - 1\}\|\psi_i\psi_i^\top\|_2 \leq \frac{1}{p_i}\psi_i^\top\psi_i.$$

However, the RHS of the above bound can be arbitrary large if p_i goes to zero. By using (30) and (31), we show that the sampling probability p_i is lower bounded as follows

$$p_i \geq \min\{c\psi_i^\top\psi_i, 1\}. \tag{32}$$

The argument to prove (32) goes as follows. First, we recall the definition of the sampling probability at step i, as it is given in Sect. 5 by

$$p_i \triangleq \min\{1, c\tilde{l}_i\},$$

with $\tilde{l}_i \triangleq \min\{1, (1+t)s_i\}$ and

$$s_i \triangleq b_i^\top (BS_{i-1}S_{i-1}^\top B^\top + \epsilon\mathbb{I})^{-1}b_i = \frac{1}{\epsilon}\left[P_{n\gamma}(K) - L_{\epsilon,S_{i-1}}(P_{n\gamma}(K))\right]_{ii},$$

while b_i is the i-th column of B. Next, we obtain the following set of inequalities

$$\tilde{l}_i = \min\{1, (1+t)b_i^\top(BS_{i-1}S_{i-1}^\top B^\top + \epsilon\mathbb{I})^{-1}b_i\}$$

$$\geq \min\{1, (1+t)b_i^\top\left(B_{[i-1]}B_{[i-1]}^\top + \epsilon\mathbb{I} + t(BB^\top + \epsilon\mathbb{I})\right)^{-1}b_i\} \quad \text{(by using(30))}$$

$$\geq \min\{1, (1+t)b_i^\top\left(BB^\top + \epsilon\mathbb{I} + t(BB^\top + \epsilon\mathbb{I})\right)^{-1}b_i\} \quad \text{(by using(31))}$$

$$\geq \min\{1, b_i^\top(BB^\top + \epsilon\mathbb{I})^{-1}b_i\} = b_i^\top(BB^\top + \epsilon\mathbb{I})^{-1}b_i = \psi_i^\top\psi_i.$$

So, we have $p_i \geq \min\{c\psi_i^\top\psi_i, 1\}$. If $p_i = 1$, then $X_i = 0$ as a direct consequence of the definition in (29), so that the step i does not change the value of the martingale. Otherwise, if $p_i < 1$, we have $p_i \geq c\psi_i^\top\psi_i$.

Concentration bound We can now estimate the probability of failure by using a concentration bound given in Theorem 3. Let us check that the hypotheses of Theorem 3 are satisfied. Clearly, each increment is bounded as follows:

$$\|X_i\|_2 \leq \max\{1, 1/p_i - 1\}\|\psi_i\psi_i^\top\|_2 \leq \frac{1}{p_i}\psi_i^\top\psi_i \leq 1/c \triangleq U,$$

almost surely. Similarly, by using once more $\psi_i^\top\psi_i \leq p_i/c$, we find:

$$\mathbb{E}_{i-1}(X_i^2) \leq p_i(1/p_i - 1)^2(\psi_i\psi_i^\top)^2 + (1-p_i)(\psi_i\psi_i^\top)^2 \leq \frac{1}{p_i}(\psi_i\psi_i^\top)^2 \leq \psi_i\psi_i^\top/c.$$

Then, the predictable quadratic variation satisfies the following upper bound

$$W_n \preceq \frac{1}{c}\sum_{i=1}^n \psi_i\psi_i^\top = \frac{1}{c}P_\epsilon(P_{n\gamma}(K)) = \frac{1}{c}\frac{1}{1+\epsilon}P_{\frac{\epsilon n\gamma}{1+\epsilon}}(K) \text{ almost surely,}$$
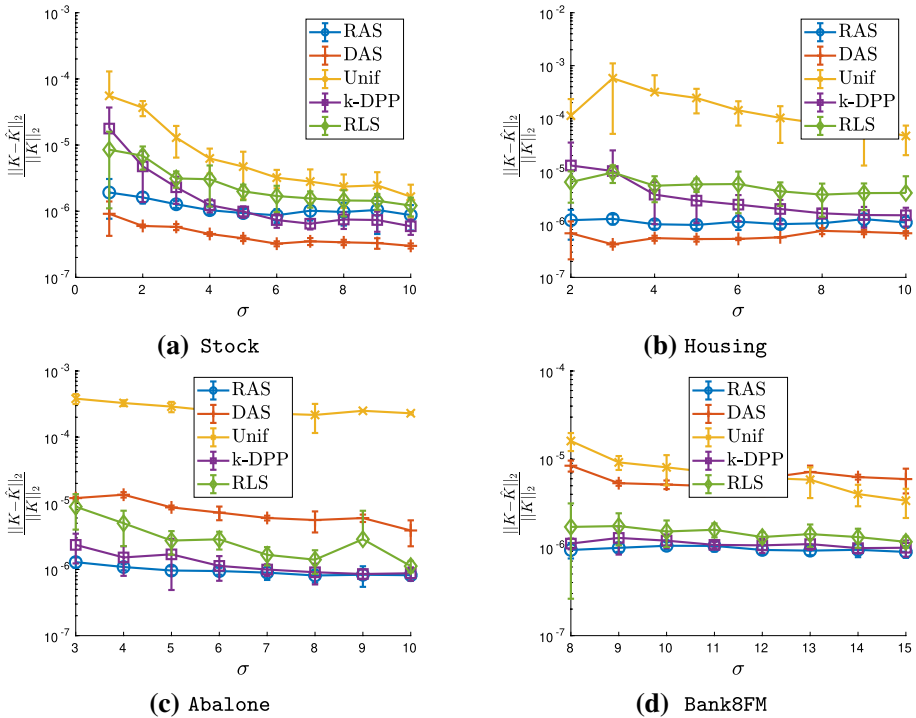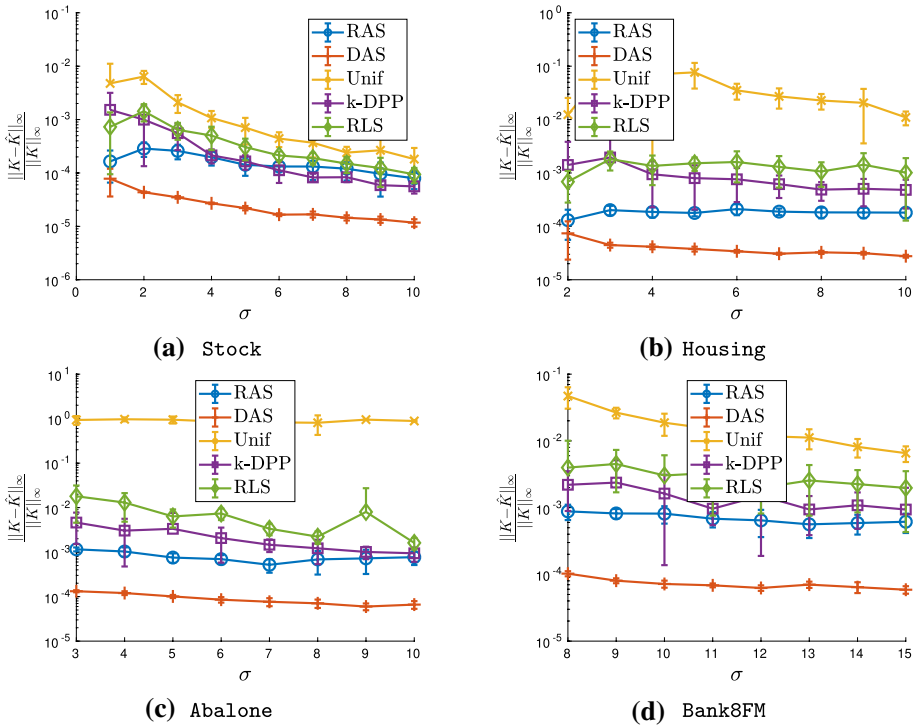
Fig. 8 The Relative operator norm of the Nyström approximation error as a function of the bandwidth σ. The error bars are the standard deviations

where the last equality is due to Lemma 2. Furthermore, we have $\lambda_{max}(W_n) \leq 1/c = \sigma^2$ almost surely[11]. Now, for a symmetric matrix M, we recall that $\min\{M, 1\}$ is defined on the eigenvalues of M. Next, we use that $\min\{\lambda, 1\} \leq \lambda$ for all real λ. Hence,

$$\text{Tr}\left(\min\{\frac{t}{U}\frac{\mathbb{E}(W_n)}{\sigma^2}, 1\}\right) \leq \frac{tc}{1+\epsilon}\text{Tr}\left(P_{\frac{c n \gamma}{1+\epsilon}}(K)\right).$$

Theorem 3 can now be used to find how large the oversampling parameter $c > 0$ has to be so that $\lambda_{\max}(Y_n) \leq t$.

Finding a lower bound on c In view of Theorem 3, we require that $t \geq \frac{1}{c}\frac{1+\sqrt{37}}{6}$. Let the accuracy parameter be $t = 1/2$ in order to simplify the result; cfr. Musco and Musco (2017) where a similar choice is done. This implies then that $c \geq \frac{1+\sqrt{37}}{3}$. Another lower bound on c is obtained by bounding the failure probability. Namely, the failure probability given by (28) in Theorem 3 is upper bounded as follows:

$$\Pr(\|Y_n\|_2 > t) \leq 50\,\text{Tr}\left(\min\{\frac{t}{U}\frac{\mathbb{E}(W_n)}{\sigma^2}, 1\}\right)e^{\frac{-t^2/2}{\sigma^2+tU/3}} \leq \frac{50tc}{1+\epsilon}d_{\text{eff}}\,e^{\frac{-t^2/2}{1/c+t/(3c)}},$$

---

[11] Notice that σ^2 denotes here an upper bound on the predictable quadratic variation and not the bandwidth of a kernel.

Fig. 9 The relative max norm of the Nyström approximation error as a function of the bandwidth σ. The error bars are the standard deviations

Table 3 Parameters for the samplings with BLESS

Data set	σ	q_0	c_0	c_1	c_2
Super	3	2	2	3	3
CASP	1	2	2	3	3
Adult	10	2	2	3	2.2
MiniBooNE	5,10	2	2	3	3.6,4
cod-RNA	4	2	2	3	2.9
Covertype	10	2	2	3	3

where $d_{\text{eff}} = d_{\text{eff}}(\epsilon n \gamma / (1 + \epsilon))$. We now find how $c > 0$ can be chosen such that this upper bound on the failure probability is smaller than δ. For $t = 1/2$ we have the condition:

$$\frac{25c}{1 + \epsilon} \frac{d_{\text{eff}}}{\delta} \leq e^{\frac{c}{8(1+1/6)}}.$$

Hence, by a change of variables, we can phrase this inequality in the form $\exp(x) \geq ax$, which can be written equivalently as follows:

$$x \geq -W_{-1}(-1/a), \qquad \text{for } x \geq 1 \text{ and } a \geq e,$$
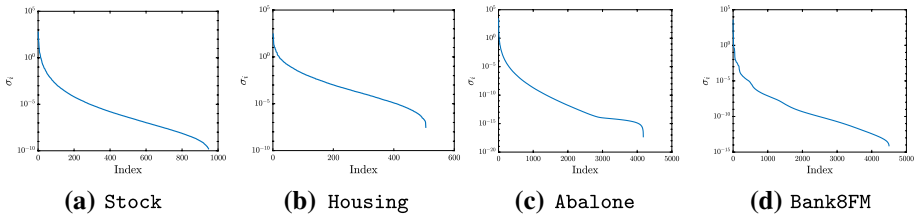
Fig. 10 Singular value spectrum of the data sets of Table 1 on a logarithmic scale. For a given index (e.g., 100), the value of the eigenvalues for `Stock` and `Housing` are considerably smaller than `Abalone` and `Bank8FM`

where $W_{-1} : [-1/e, 0(\to \mathbb{R}$ is the negative branch of the Lambert function (see Section 3. of Corless et al. (1996)), which is a monotone decreasing function. Notice that the asymptotic expansion of this branch of the Lambert function is $\lim_{y \to 0_-} W_{-1}(y)/\log(-y) = 1$.
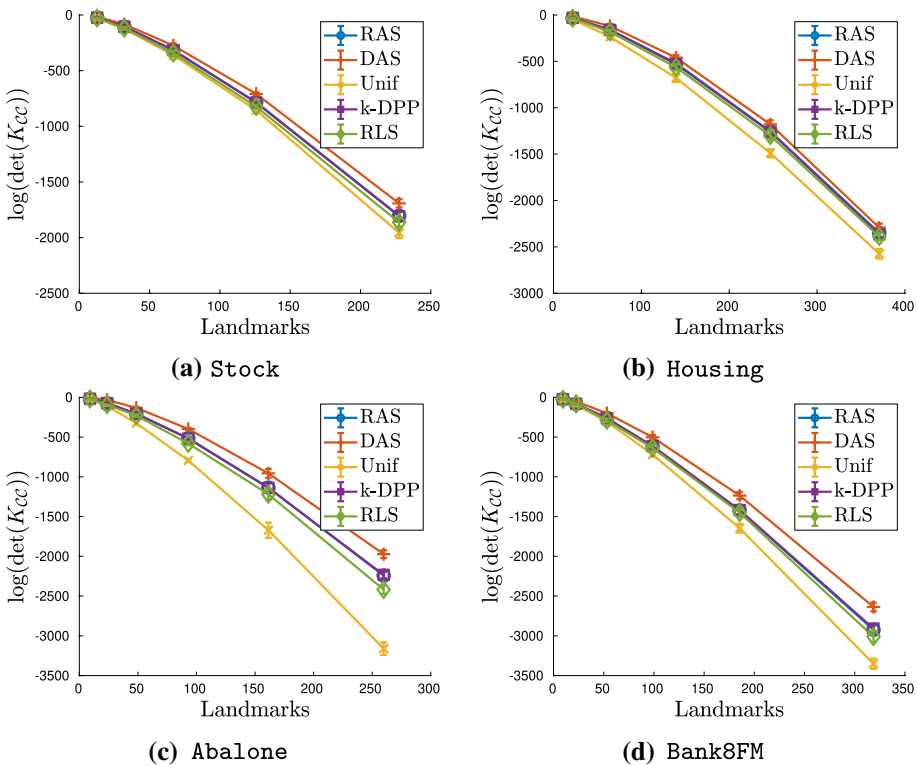
Fig. 11 Diversity as a function of the number of landmarks. The $\log\det(K_{CC})$ is plotted on a logarithmic scale, averaged over 10 trials. The error bars are the standard deviations

Therefore, we can write $-W_{-1}(-1/a) \sim \log(a)$ (with $a > 0$), so that we finally have the conditions:

$$c \geq -\frac{28}{3} W_{-1}\left(-\frac{3(1+\epsilon)\delta}{700 d_{\text{eff}}}\right), \text{ and } c \geq \frac{1+\sqrt{37}}{3},$$

where $d_{\text{eff}} = d_{\text{eff}}(\epsilon n \gamma/(1+\epsilon))$. This proves the desired result. □

Appendix E: Ablation study for different bandwidths of the Gaussian kernel

Here we repeat the Nyström experiment of Sect. 6.1 for different bandwidths of the Gaussian kernel. Namely, we take a fixed regularization $\gamma = 10^{-4}$ in RAS, but vary the bandwidth σ of the Gaussian kernel. The parameter c is not changed and remains the same as in Table 1 for all the datasets. Note that a different σ results in a different spectrum of the kernel matrix, consequently the RAS algorithm outputs a different number of landmarks for each setting. The best performing subset is chosen as representative. The experiment is repeated 10 times and the averaged results are given on Figs. 8 and 9. The conclusions from the experiment in Sect. 6.1 are shown to be consistent over a range of σ's.

(a) Stock

(b) Housing

(c) Abalone

(d) Bank8FM
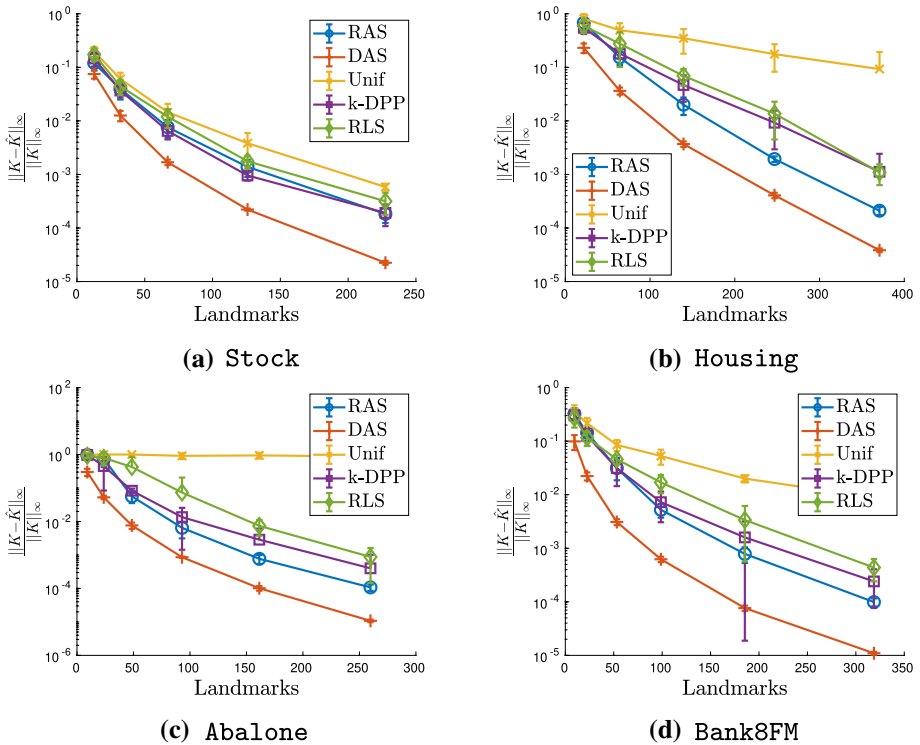
Fig. 12 Relative max norm with error bars of the approximation as a function of the number of landmarks. The error is plotted on a logarithmic scale, averaged over 10 trials

(a) Stock

(b) Housing

(c) Abalone

(d) Bank8FM
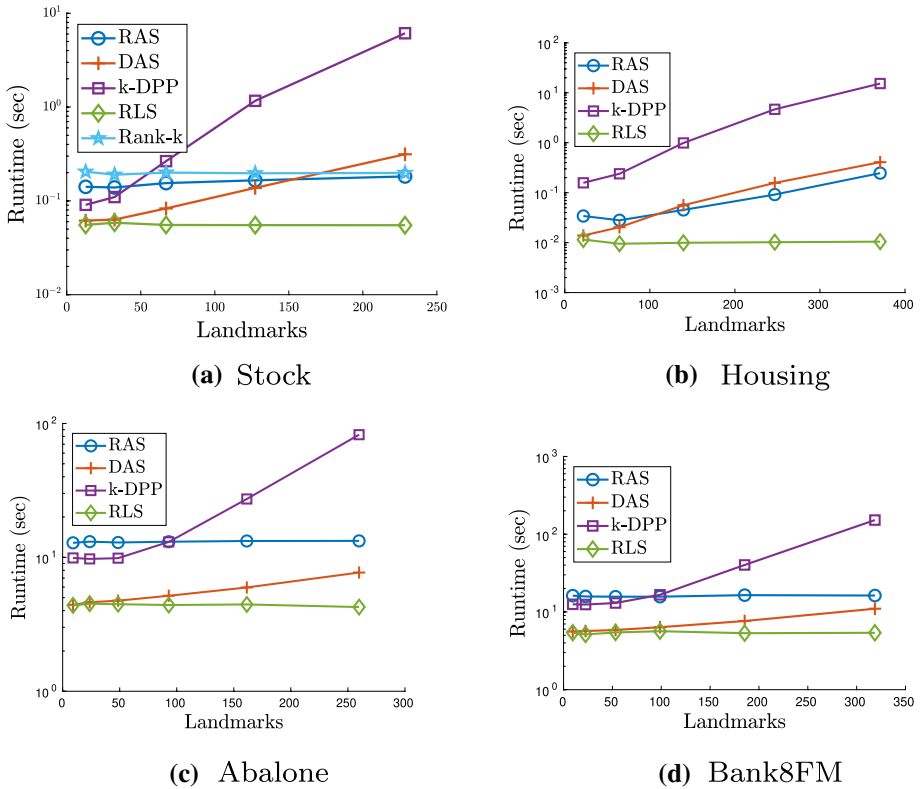
Fig. 13 Timings for the computations of Fig. 3 as a function of the number of landmarks. The timings are plotted on a logarithmic scale, averaged over 10 trials

Appendix F: Simulation details, additional tables and figures

The computer used for the small-scale simulations has a processor with 8 cores $3.40GHz$ and 15.5 GB of RAM. The implementation of the algorithms is done with matlabR2018b. For the medium scale experiments, the implementation of RAS was done in Julia1.0 while the simulations were performed on a computer with a 12×3.7 GHz processors and 62 GB RAM.

Also, we provide here additional explanatory figures concerning the data sets of Table 1. In Fig. 10, the eigenvalues of the kernel matrices are displayed. The decay of the spectrum is faster for `Stock` and `Housing`. Furthermore, the accuracy of the kernel approximation for the max norm is also given in Fig. 12. Namely, the DAS method of Algorithm 1 shows a better approximation accuracy for this specific norm. As a measure of diversity the $\text{logdet}(K_{\mathcal{CC}}) = \sum_{i=1}^{|\mathcal{C}|} \log(\lambda_i)$ with $\lambda_1, ..., \lambda_{|\mathcal{C}|}$ the eigenvalues of $K_{\mathcal{CC}}$, is given in Fig. 11. The DAS method shows the largest determinant, RAS has a similar determinant as the k-DPP (Figs. 11, 12, 13).

Data availability All data sets used in this paper are public; see the following links https://www.cs.toronto.edu/~delve/data/data sets.html, https://www.openml.org/d/223 and https://archive.ics.uci.edu/ml/datasets.php.

Declarations

References

Askari, A., Yang, F., El Ghaoui, L. (2018). *Kernel-based outlier detection using the inverse Christoffel function*. arxiv preprint arxiv:1806.06775

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In: *Proceedings of the COLT Conference on Learning Theory*, pp. 185–209

Belabbas, M. A., & Wolfe, P. J. (2009). Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences, 106*(2), 369–374.

Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., & Wojtaszczyk, P. (2011). Convergence rates for greedy algorithms in reduced basis methods. *SIAM Journal on Mathematical Analysis, 43*(3), 1457–1472.

Calandriello, D., Lazaric, A., & Valko, M. (2017a). Distributed adaptive sampling for kernel matrix approximation. In: Singh A, Zhu XJ (eds) *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 20–22 April 2017, Fort Lauderdale, FL, USA, PMLR, Proceedings of Machine Learning Research, vol. 54, pp. 1421–1429.

Calandriello, D., Lazaric, A., & Valko, M. (2017b). Second-order kernel online convex optimization with adaptive sketching. In: *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, JMLR.org, pp. 645–653.

Chen, V., Wu, S., Ratner, A. J., Weng, J., & Ré, C. (2019). Slice-based learning: A programming model for residual learning in critical data slices. In: *Advances in neural information processing systems*, pp. 9392–9402.

Cohen, M. B., Musco, C., & Pachocki, J. (2015). Online row sampling. In: *Proceedings of the 19th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., & Knuth, D. E. (1996) On the Lambert W Function. In: *Advances in computational mathematics*, pp. 329–359.

Cortes C, Mohri M, Talwalkar A (2010) On the impact of kernel approximation on learning accuracy. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 113–120.

Derezinski, M., Calandriello, D., & Valko, M. (2019). Exact sampling of determinantal point processes with sublinear time preprocessing. In: *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Derezinski, M., Khanna, R., & Mahoney, M. (2020). Improved guarantees and a multiple-descent curve for the column subset selection problem and the nyström method. *To appear at NeurIPS 2020 abs/2002.09073*.

DeVore, R., Petrova, G., & Wojtaszczyk, P. (2013). Greedy algorithms for reduced bases in Banach spaces. *Constructive Approximation, 37*(3), 455–466.

Drineas, P., & Mahoney, M. W. (2005). On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research, 6,* 2153–2175.

El Alaoui, A., & Mahoney, M. (2015). Fast randomized kernel ridge regression with statistical guarantees. *Advances in Neural Information Processing Systems, 28,* 775–783.

Fanuel, M., Schreurs, J., & Suykens, J. (2021). Diversity sampling is an implicit regularization for kernel methods. *SIAM Journal on Mathematics of Data Science, 3*(1), 280–297.

Farahat, A., Ghodsi, A., & Kamel, M. (2011). A novel greedy algorithm for Nyström approximation. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research,*, vol. 15, pp. 269–277.

Feldman, V. (2020). Does learning require memorization? a short tale about a long tail. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959

Gao, T., Kovalsky, S., & Daubechies, I. (2019). Gaussian process landmarking on manifolds. *SIAM Journal on Mathematics of Data Science, 1*(1), 208–236.

Gartrell, M., Brunel, V. E., Dohmatob, E., & Krichene, S. (2019). Learning nonsymmetric determinantal point processes. *Advances in Neural Information Processing Systems, Curran Associates Inc, 32,* 6718–6728.

Gauthier, B., & Suykens, J. (2018). Optimal quadrature-sparsification for integral operator approximation. *SIAM Journal on Scientific Computing, 40*(5), A3636–A3674.

Girolami, M. (2002). Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computing, 14*(3), 669–688.

Gittens, A., & Mahoney, M. W. (2016). Revisiting the Nyström method for improved large-scale machine learning. *Journal of Machine Learning Research,17,* 117:1–11765.

Gong, B., Chao, W. L., Grauman, K., Sha, F. (2014). Diverse sequential subset selection for supervised video summarization. In: *Advances in Neural Information Processing Systems*, pp. 2069–2077.

Guyon, I., Matic, N., & Vapnik, V. (1996), Advances in knowledge discovery and data mining. *Chap Discovering Informative Patterns and Data Cleaning*, pp. 181–203.

Hough, J. B., Krishnapur, M., Peres, Y., & Virág, B. (2006). Determinantal processes and independence. *Probability Surveys, 3,* 206–229.

Kulesza, A., & Taskar, B. (2010), Structured determinantal point processes. In: *Advances in neural information processing systems*, pp. 1171–1179,

Kulesza, A., & Taskar, B. (2011), k-dpps: Fixed-size determinantal point processes. In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1193–1200.

Kulesza, A., & Taskar, B. (2012a). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning, 5*(2–3), 123–286.

Kulesza, A., & Taskar, B. (2012b) Learning determinantal point processes. *arxiv preprint*arxiv:1202.3738.

Lasserre, J., & Pauwels, E. (2017), The empirical Christoffel function with applications in Machine Learning, *Arxiv preprint*arXiv:1701.02886.

Launay, C., Galerne, B., & Desolneux, A. (2020). Exact sampling of determinantal point processes without eigendecomposition. *Journal of Applied Probability, 57*(4), 1198–1221.

Liang, D., & Paisley, J. (2015), Landmarking manifolds with gaussian processes. In: *Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol. 37, pp. 466–474.

Li, C., Jegelka, S,, & Sra, S. (2016a) Efficient sampling for k-determinantal point processes. In: *AISTAT*.

Li, C., Jegelka, S,, & Sra, S. (2016b). Fast Dpp. sampling for Nyström with application to kernel methods. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning*—Volume 48, ICML'16, pp. 2061–2070.

McCurdy, S. (2018). Ridge regression and provable deterministic ridge leverage score sampling. *Advances in Neural Information Processing Systems, 31,* 2468–2477.

Minsker, S. (2017). On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters, 127,* 111–119.

Musco, C., & Musco, C. (2017). Recursive sampling for the Nyström method. *Advances in Neural Information Processing Systems, 30,* 3833–3845.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020), Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159,

Paisley, J., Liao, X., & Carin, L. (2010). Active learning and basis selection for kernel-based linear models: A bayesian perspective. *IEEE Transactions on Signal Processing, 58*(5), 2686–2700.

Papailiopoulos, D., Kyrillidis, A., & Boutsidis, C. (2014). Provable deterministic leverage score sampling. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, KDD '14, pp. 997–1006.

Pauwels, E., Bach, F., & Vert, J. (2018). Relating leverage scores and density using regularized Christoffel functions. In: *Advances in Neural Information Processing Systems, 32,* 1663–1672.

Pinkus, A. (1979). Matrices and *n*-widths. *Linear Algebra and Applications, 27,* 245–278.

Poulson, J. (2020). High-performance sampling of generic determinantal point processes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 378*(2166), 20190059.

Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1177–1184.

Rasmussen, C., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. Cambridge: The MIT Press.

Rudi, A., Calandriello, D., Carratino, L., & Rosasco, L. (2018). On fast leverage score sampling and optimal learning. *Advances in Neural Information Processing Systems, 31,* 5677–5687.

Rudi, A., Carratino, L., & Rosasco, L. (2017). Falkon: An optimal large scale kernel method. *Advances in Neural Information Processing Systems, 30,* 3888–3898.

Rudi, A., De Vito, E., Verri, A., & Odone, F. (2017). Regularized kernel algorithms for support estimation. *Frontiers in Applied Mathematics and Statistics, 3,* 23.

Santin, G., & Haasdonk, B. (2017). Convergence rate of the data-independent p-greedy algorithm in kernel-based approximation. *Journal Dolomites Research Notes on Approximation, 10,* 68–78.

Schreurs, J., Fanuel, M., & Suykens, J. (2020). Ensemble kernel methods, implicit regularization and determinantal point processes. *ICML 2020 workshop on Negative Dependence and Submodularity, PMLR 119*.

Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., & Vandewalle, J. (2002). *Least Squares Support Vector Machines*. Singapore: World Scientific.

Tremblay, N., Barthelmé, S., & Amblard, P. (2019). Determinantal point processes for coresets. *Journal of Machine Learning Research, 20*(168), 1–70.

Tropp, J. (2019), Matrix concentration & computational linear algebra. *Teaching Resource* (Unpublished).

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability, 16,* 262–270.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE, 9*(1), e84217.

Williams, C., & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems, 13,* 682–688.

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