DIVERSITY SAMPLING IS AN IMPLICIT REGULARIZATION FOR KERNEL METHODS *

MICHAËL FANUEL†, JOACHIM SCHREURS†, AND JOHAN A.K. SUYKENS†

Abstract. Kernel methods have achieved very good performance on large scale regression and classification problems, by using the Nyström method and preconditioning techniques. The Nyström approximation – based on a subset of landmarks – gives a low rank approximation of the kernel matrix, and is known to provide a form of implicit regularization. We further elaborate on the impact of sampling *diverse* landmarks for constructing the Nyström approximation in supervised as well as unsupervised kernel methods. By using Determinantal Point Processes for sampling, we obtain additional theoretical results concerning the interplay between diversity and regularization. Empirically, we demonstrate the advantages of training kernel methods based on subsets made of diverse points. In particular, if the dataset has a dense bulk and a sparser tail, we show that Nyström kernel regression with diverse landmarks increases the accuracy of the regression in sparser regions of the dataset, with respect to a uniform landmark sampling. A greedy heuristic is also proposed to select diverse samples of significant size within large datasets when exact DPP sampling is not practically feasible.

1. Introduction. Kernel methods often rely on low rank approximations to deal with large scale datasets. This paper addresses the special case of Nyström approximation that is defined hereafter. Namely, let $k(x,y) > 0$ be a continuous and strictly positive definite kernel. Given data $\{x_i \in \mathbb{R}^d\}_{i\in[n]}$, kernel methods rely on the entries of the Gram matrix $K = [k(x_i,x_j)]_{i,j}$. To deal with large scale problems, one often samples a subset of landmarks $\mathcal{C} \subseteq [n]$ and defines a $n \times |\mathcal{C}|$ sampling matrix C obtained by selecting the columns of the identity matrix indexed by $\mathcal{C}$. This is useful to select rectangular and squared submatrices as follows: $K_{\mathcal{C}} = KC$ and $K_{\mathcal{C}\mathcal{C}} = C^\top KC$. Then, the $n \times n$ kernel matrix K is approximated by a low rank Nyström approximation

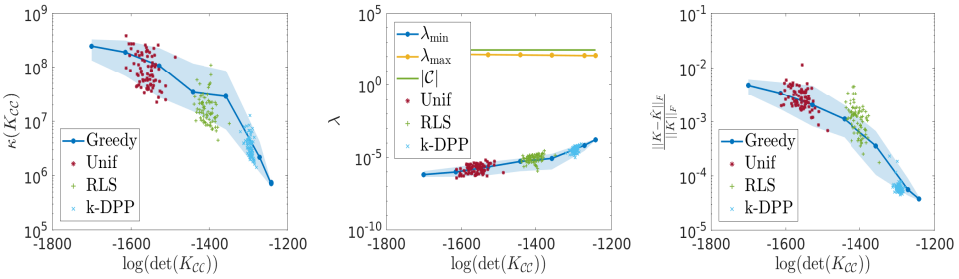
$$\hat{K}(K,\mathcal{C}) = K_{\mathcal{C}} K_{\mathcal{C}\mathcal{C}}^{-1} K_{\mathcal{C}}^\top, \tag{1.1}$$

which involves the inverse of $K_{\mathcal{C}\mathcal{C}}$. However, this submatrix can be ill-conditioned. In practice, this happens especially when $\mathcal{C}$ is sampled uniformly at random and for the Gaussian kernel $k(x,y) = \exp(-\|x-y\|_2^2/\sigma^2)$, that is used in this paper. We argue here that a sampling of $\mathcal{C}$ which yields a good performance is closely related to the *diversity* of the set of landmarks. In this work, the diversity of $\mathcal{C}$ is measured by the value of $\det(K_{\mathcal{C}\mathcal{C}})$. This is intuitively understood thanks to the connection between determinant and volume [15]. We claim that selecting a diverse sample *implicitly regularizes* the corresponding submatrix. This is illustrated on the `Housing`

* **Funding:** * EU: The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.Research Council KUL: Optimization frameworks for deep kernel machines C14/18/068. Flemish Government:FWO: projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations), PhD/Postdoc grant.Impulsfonds AI: VR 2019 2203 DOC.0318/1QUATER Kenniscentrum Data en Maatschappij. Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders(FWO) and the Flemish Government – department EWI.

†KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. email: michael.fanuel@kuleuven.be

dataset in Figure 1 where the Nyström approximation error and condition number of $K_{\mathcal{C}\mathcal{C}}$ are given for several $\mathcal{C}$ of identical cardinality, associated with different diversities. Additional technical details are given in Section 3. The empirical findings described in Section 3 indicate that Nyström-based kernel methods are improved if the landmarks are both diverse and yield an accurate kernel approximation. To illustrate this statement, we use a greedy swapping algorithm, namely Algorithm 2.1 (blue line in Figure 1), which allows for obtaining a sample of landmarks with a given diversity. It is worth mentioning that different other methods exist to sample diverse landmarks, such as volume sampling [8], greedy methods [7], Determinantal Point Processes (DPP) [15], etc. The example of Figure 1 illustrates the connection between diversity, regularization and Nyström approximation error. Namely, we sample repeatedly subsets of the same size by using Uniform, Ridge Leverage Score (RLS) [9] and DPP sampling (defined hereafter), which yield samples with an increasing diversity. Figure 1 shows that the corresponding kernel submatrices have an increasing least eigenvalue, a decreasing Nyström approximation error and a decreasing condition number. This highlights the implicit regularization due to diversity. Among diverse sampling methods, DPPs provide a natural probabilistic framework



(a) $\kappa(K_{\mathcal{C}\mathcal{C}})$ (b) $\lambda_{\min}(K_{\mathcal{C}\mathcal{C}})$ (c) Approximation error

Fig. 1: Nyström approximation for the `Housing` dataset with parameters given in Table 1. The condition number, smallest and largest eigenvalues of $K_{\mathcal{C}\mathcal{C}}$, relative Frobenius norm of the approximation error of $\hat{K} = \hat{K}(K, \mathcal{C})$ are plotted versus the sample diversity. The larger $\det(K_{\mathcal{C}\mathcal{C}})$, the more diverse the subset. Error bars are the 0.05 and 0.95 quantile over 10 simulations.

for diversity sampling. Their elegant definition allows to derive results formalizing our empirical observations about the interplay between regularization and diversity. Let us briefly define them in the simplest setting, while a more complete overview can be found in [15].

1.1. L-ensemble DPP sampling. Let L be a $n \times n$ positive definite symmetric matrix. An L-ensemble, denoted by $DPP_L(L)$, is a DPP with a specific probability mass function for $\mathcal{C} \subseteq [n]$, which depends on the matrix L. Namely, when $\mathcal{Y} \sim DPP_L(L)$, we have

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

where the normalization is designed such that these probabilities sum up to 1 (see [15]). In this paper, we define $L = K/\alpha$ where $\alpha > 0$ is a scale parameter that influences the expected subset size. Classically, an alternative viewpoint deals with the inclusion

probabilities as given by $\Pr(\mathcal{C} \subseteq \mathcal{Y}) = \det(P_{\mathcal{C}\mathcal{C}})$, where

$$P = K(K + \alpha\mathbb{I})^{-1}, \tag{1.2}$$

is the marginal kernel corresponding to the L-ensemble with $L = K/\alpha$. The diagonal of the soft projector matrix (1.2) yields the so-called Ridge Leverage Scores (RLS) of the data points:

$$\ell_i = P_{ii} \text{ for } i \in [n],$$

where ℓ_i denotes the i-th entry of the vector $\boldsymbol{\ell} \in \mathbb{R}^n$ containing the RLS. These scores have been used in order to sample landmarks points in various works [18, 9, 1] in the context of Nyström approximations. RLS can be considered as a measure of importance or 'outlierness' of a data point. Moreover, connections between RLS, DPPs and Christoffel functions were explored in [10]. The sum of the RLS yields the effective dimension $d_{\text{eff}}(K/\alpha)$ which is also the expected size $d_{\text{eff}}(K/\alpha) = \mathbb{E}_{\mathcal{C}}[|\mathcal{C}|]$ if $\mathcal{C} \sim DPP_L(K/\alpha)$. Since the subset size $|\mathcal{C}|$ in itself is also a random variable, it is also customary to use k-DPPs which are DPPs conditioned on a given subset size k (see also [16]).

The following two sections motivate the impact of the regularity of $K_{\mathcal{C}\mathcal{C}}$ in two applications. Firstly, a better kernel approximation yields an improvement of the performance of unsupervised kernel methods such as Kernel Principal Component Analysis (KPCA) [24, 25] and Kernel k-means [27]. Secondly, the conditioning of $K_{\mathcal{C}\mathcal{C}}$ is also important for large-scale supervised learning methods – based on Nyström approximation – as the convergence and accuracy of iterative solvers depends often of the condition number. Finally, sampling with a diverse method spreads the points more over the full dataset. This is especially important for accuracy in less populated or 'outlying' regions in the dataset.

We now give a short overview of how the Nyström approximation is used to speed up kernel PCA and kernel ridge regression.

1.2. Kernel PCA. The Nyström method is used to develop a more computationally efficient approximate kernel PCA algorithm [25]. Let $\mathcal{H}$ be the Reproducing Kernel Hilbert Space associated to k and assume that the data is sampled from a distribution ρ such that $\mathbb{E}_{X\sim\rho}[f(X)] = 0$ for all $f \in \mathcal{H}$. We recall that Kernel PCA is a principal component analysis in a RKHS, i.e., it consists in finding the directions of maximum variance. Indeed, let $k_{x_i}(\cdot) = k(x_i, \cdot)$ and let the empirical covariance operator be $C = \frac{1}{n}\sum_{i=1}^n k_{x_i} \otimes k_{x_i}$. Also, we define the subspace $\mathcal{H}_{\mathcal{C}} = \text{span}\{k_{x_i} \text{ s.t. } i \in \mathcal{C}\}$. Then, the optimization problem

$$\sup_f \langle f, Cf\rangle_{\mathcal{H}} \text{ s.t. } \|f\|_{\mathcal{H}} = 1 \text{ and } f \in \mathcal{H}_{\mathcal{C}},$$

corresponds to a Nyström approximation of KPCA if $\mathcal{C} \subset [n]$. The empirical estimation of KPCA involves the eigendecomposition of the matrix

$$M = K_{\mathcal{C}\mathcal{C}}^{-1/2} K_{\mathcal{C}}^\top K_{\mathcal{C}} K_{\mathcal{C}\mathcal{C}}^{-1/2},$$

sharing its non-zero eigenvalues with (1.1). Let $(\hat{\lambda}_{\ell,\mathcal{C}}, \boldsymbol{u}_\ell)_{\ell=1}^{|\mathcal{C}|}$ be eigenpairs of M sorted in descending order. KPCA aims to construct the orthogonal projector $P_{\mathcal{C}} = \sum_{\ell=1}^c \boldsymbol{u}_\ell \boldsymbol{u}_\ell^\top$ on the subspace corresponding to the c largest eigenvalues, with $c \leq |\mathcal{C}|$. Clearly, the conditioning of $K_{\mathcal{C}\mathcal{C}}$ is important for this task in view of the definition of

M. The reconstruction error for c components assesses the quality of the approximation and is given by:

$$\hat{R}\left(P_{\mathcal{C}}\right) = \operatorname{tr}(K) - \sum_{\ell=1}^{c} \boldsymbol{u}_\ell^\top M \boldsymbol{u}_\ell = \sum_{\ell=1}^{n} \hat{\lambda}_\ell - \sum_{\ell=1}^{c} \hat{\lambda}_{\ell,\mathcal{C}}, \tag{1.3}$$

where $(\hat{\lambda}_\ell, \boldsymbol{v}_\ell)_{\ell=1}^n$ are eigenpairs of K. A small reconstruction error is then achieved thanks to an accurate Nyström approximation as detailed in Section 2.2.

1.3. Regression. In approximate Kernel Ridge Regression (KRR), the regressor is obtained from given input-output pairs $\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i\in[n]}$ by solving

$$f^\star = \arg\min_{f\in\mathcal{H}_\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \gamma\|f\|_{\mathcal{H}}^2, \text{ with } \gamma > 0, \tag{1.4}$$

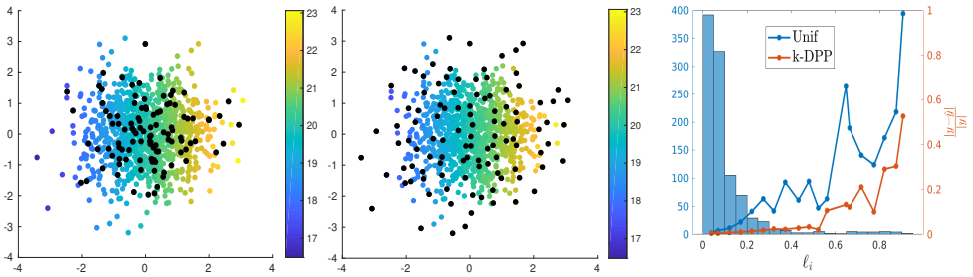
where $\mathcal{H}_\mathcal{C} = \operatorname{span}\{k(x_i,\cdot)|i \in \mathcal{C}\}$ and $\mathcal{C} \subseteq [n]$. The regressor is $f^\star(\cdot) = \sum_{i\in\mathcal{C}} \alpha_i^\star k(x_i,\cdot)$ with

$$\boldsymbol{\alpha}^\star = (K_{\mathcal{C}}^\top K_{\mathcal{C}} + n\gamma K_{\mathcal{C}\mathcal{C}})^{-1} K_{\mathcal{C}}^\top \boldsymbol{y}. \tag{1.5}$$

The condition number of (1.5) crucially depends on the magnitude of the least eigenvalue of $K_{\mathcal{C}\mathcal{C}}$ which plays the role of regularization term. Notice that the full KRR is simply obtained by replacing $\mathcal{H}_\mathcal{C}$ by $\mathcal{H}$ in (1.4).

Stability of the expected risk. The first consequence of an accurate Nyström approximation is that the expected risk of approximate KRR is upper bounded by a controllable constant as it is explained in Section 2.3. This means that the training problem can not be dramatically affected by the Nyström approximation.

Uniform test error. A second consequence is directly related to the diversity of the landmarks and is illustrated in Figure 2. Namely, Figure 2 shows the training dataset which consists of 1000 sampled points generated as follows: $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I})$ and $y_i = \boldsymbol{e}_1^\top \boldsymbol{x}_i + b + \epsilon_i$ with iid $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with $\sigma = 0.1$ and $b = 20$. The corresponding y_i values are visualized by the color of the points. Landmarks are then sampled by



(a) Uniform sampling. (b) k-DPP sampling. (c) RLS and binned MAPE.

Fig. 2: Toy example of regression. In 2c, a histogram of RLS distribution and the MAPE test error in each bin are displayed.

using uniform and DPP sampling. Uniform sampling oversamples the dense parts, while a diverse sampling algorithm spreads the points over the full dataset. A kernel

ridge regressor using a Gaussian kernel with $\sigma = 0.5$ is trained by using (1.4), where the optimal regularization parameter $\gamma > 0$ is determined by using cross-validation. Figure 2c shows the Mean Absolute Percentage Error (MAPE) as a function of the ridge leverage scores of the test set, where each dot corresponds to the MAPE in the corresponding bin of the histogram. This stratification of the dataset allows to visualize how the regressor performs in dense (small RLS) and sparser (large RLS) groups of the dataset.

Diverse sampling has consistently a better performance with respect to uniform sampling, while the difference is more apparent for high leverage scores. This is especially important when sampling from datasets with long tail RLS distributions. Hence, in the case of diverse sampling, we emphasize that the percentage error is more uniform on the support of the dataset while the regressor makes a smaller error on points with larger leverage score compared to regressors obtained with uniform sampling, while the total MAPE shows only a minor difference. Additional illustrations of this effect are given in Section 3. Naturally, diverse sampling is less important if there is no long tail of in the RLS distribution.

We now want to emphasize why diverse sampling is important, especially in stratified datasets. Recently, there has been a lot of interest in not only predicting well in the majority of the data, but also for specific outlying points [26, 20, 6]. These outlying points can, e.g., correspond to serious diseases in a medical dataset, being less common than mild diseases. Incorrectly classifying these outliers could lead to significant harm to patients. 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, obscuring the fact that there may be an unidentified subset of cases where the performance is poor. These stratifications often occur in datasets with a long tail, i.e. the data distribution of each class is viewed as a mixture of distinct subpopulations [11]. For example, images of dogs include different species photographed from different perspectives and under different conditions (such as close-ups, in the woods and during the rain). A long-tailed mixture distribution will have some subpopulations from which just a few or only a single one example was observed. When using sampling algorithms, it is therefore necessary to select points out of each subpopulation to achieve close-to-optimal generalization error. One expects that, before seeing the dataset, the learning algorithm does not know the frequencies of subpopulations and may not be able to predict accurately on a subpopulation without observing any examples from it. By making sure the selected subset is *diverse* enough, there is a higher chance of every subpopulation being included in the sample. In [11], it is argued that datasets with long tails are a possible reason why interpolating models or models that achieve zero error rate on the training data, can still generalize well [5, 17, 4]. These hidden stratifications motivate the search for better loss functions. We therefore propose an unsupervised approach, where the loss function is determined on two parts of the data: the bulk and tail of the data. The bulk and tail of the data correspond to points with low and high outlyingness respectively, where the outlyingness is measured by the ridge leverage scores. By splitting the loss function into two parts, one can identify if the model is not only focusing on the majority data but also performing well in 'outlying' subpopulations.

The rest of the paper includes theoretical results in Section 2, and numerical experiments in Section 3. The proofs and dataset description are given in appendix. Another application, namely kernel k-means, and additional numerical experiments can be found in supplementary material.

2. Main results: Implicit regularization. On expectation, the largest and smallest eigenvalues of several matrices obtained by DPP sampling can be bounded, showing indeed that the spectrum of those submatrices are likely to be under control. This is formalized[1] in Theorem 2.1, where we denoted by $\circ$ the entry-wise product between matrices.

THEOREM 2.1 (Implicit regularization). *Let $\mathcal{C} \sim DPP_L(K/\alpha)$ and let C be a sampling matrix associated to the set $\mathcal{C}$. Then, we have*

$$\mathbb{E}_C\left[CK_{\mathcal{C}\mathcal{C}}^{-1}C^\top\right] = (K+\alpha\mathbb{I})^{-1}, \text{ and } \mathbb{E}_C\left[CK_{\mathcal{C}\mathcal{C}}C^\top\right] = P_{(2)} \circ K,$$

with $P_{(2)} = \mathrm{Diag}(\boldsymbol{\ell}) + \boldsymbol{\ell}\boldsymbol{\ell}^\top - P \circ P \succeq 0$.

It is worth remarking that the strength of the expected implicit regularization originates from the scaling of matrix defining the L-ensemble $DPP_L(K/\alpha)$. Also, notice that $P_{(2)}$ is positive semi-definite. Furthermore, if $\mathrm{diag}(K) = \mathbf{1}$ as in the case of the Gaussian kernel, the largest eigenvalue

$$\lambda_{\max}(\mathbb{E}_C\left[CK_{\mathcal{C}\mathcal{C}}C^\top\right]) \leq \lambda_{\max}(P_{(2)}) \leq \|\boldsymbol{\ell}\|_\infty + \|\boldsymbol{\ell}\|_2^2, \tag{2.1}$$

is bounded in terms of the leverage scores. This is a direct consequence of a Corollary 2 in [2], namely the spectrum of $A \circ B$ is majorized by the spectrum of A, if A and B are symmetric and positive semidefinite with $\mathrm{diag}(B) = \mathbf{1}$. It is noticeable that the largest eigenvalue of the expected kernel submatrix is under control for DPP sampling. Indeed, other sampling schemes are not known to yield similar guarantees. Again, if $\mathrm{diag}(K) = \mathbf{1}$, the trace of $\mathbb{E}_{\mathcal{C}}\left[C^\top K_{\mathcal{C}\mathcal{C}}C\right])$ is the expected size of the sample, $d_{\mathrm{eff}}(K/\alpha)$, which gives then an upper bound for $\lambda_{\max}(\mathbb{E}_C\left[CK_{\mathcal{C}\mathcal{C}}C^\top\right])$. We observe empirically that the latter yields a much larger upper bound compared to (2.1). Importantly, the scale parameter $\alpha > 0$ both controls the size of the sample and regularizes the subkernels matrix in the following sense:

$$\lambda_{\max}(\mathbb{E}_{\mathcal{C}}\left[CK_{\mathcal{C}\mathcal{C}}^{-1}C^\top\right]) \leq \alpha^{-1},$$

where $\mathcal{C} \sim DPP_L(K/\alpha)$. These results on expectation can be instructive since we expect concentration about the mean. Permantle and Peres showed that strong Rayleigh measures – generalizing DPPs – obey Gauss-Poisson concentration bounds [21]. Corollary 2.2 is then a direct consequence of that concentration result. For convenience, we write $\boldsymbol{w}_{\mathcal{C}} = C^\top \boldsymbol{w}$, where C is the sampling matrix associated to $\mathcal{C} \subseteq [n]$.

COROLLARY 2.2 (Regularization with high probability). *Let $\boldsymbol{w} \in \mathbb{R}^n$ a fixed vector such that $\|\boldsymbol{w}\|_2 = 1$. Let $\mathcal{C} \sim DPP_L(K/\alpha)$. Then, we have*

$$\left|\boldsymbol{w}_{\mathcal{C}}^\top K_{\mathcal{C}\mathcal{C}}^{-1}\boldsymbol{w}_{\mathcal{C}} - \boldsymbol{w}^\top(K+\alpha\mathbb{I})^{-1}\boldsymbol{w}\right| \leq \frac{8}{\lambda_{\min}(K)}\left(2+\sqrt{\frac{d_{\mathrm{eff}}(K/\alpha)}{2}}\right)\log\left(\frac{5}{\delta}\right),$$

with probability at least $1-\delta$, where $d_{\mathrm{eff}}(K/\alpha) = \mathrm{Tr}(K(K+\alpha)^{-1})$ is the expected subset size of $DPP_L(K/\alpha)$.

A drawback of Corollary 2.2 is that the bound hereabove depends on the inverse of $\lambda_{\min}(K)$ which may be a large number. The result may be improved by finding a better upper bound on the Lipschitz constant of the function $f(\mathcal{C}) = \boldsymbol{w}_{\mathcal{C}}^\top K_{\mathcal{C}\mathcal{C}}^{-1}\boldsymbol{w}_{\mathcal{C}}$. We refer to the proof of Corollary 2.2 for more details.

[1] After the completion of this work, a similar formula for an expected pseudo inverse was given in [19] in the context of stochatic optimization.

2.1. Nyström approximation. The subset obtained thanks to a DPP sampling is not only yielding a regular kernel submatrix, Corollary 2.3 states that it produces a good Nyström approximation. It also gives a natural connection between the projector

$$\mathbb{P}_{\mathrm{range}(K^{1/2}C)} = K^{1/2}C(K_{\mathcal{C}\mathcal{C}})^{-1}C^\top K^{1/2},$$

and the marginal kernel (1.2).

COROLLARY 2.3 (Expected Nyström approximation). *Let $\mathcal{C} \sim DPP_L(K/\alpha)$. Then, we have an expression for the Nyström error on expectation*

$$\mathbb{E}_{\mathcal{C}}\left[K - \hat{K}(K,\mathcal{C})\right] = \alpha P, \text{ and } \mathbb{E}_{\mathcal{C}}\left[\mathbb{P}_{\mathrm{range}(K^{1/2}C)}\right] = P.$$

As a straightforward consequence of Corollary 2.3, the nuclear norm of the approximation error is simply $\mathbb{E}_{\mathcal{C}}[\mathrm{Tr}(K - \hat{K}(K,\mathcal{C}))] = \alpha d_{\mathrm{eff}}(K/\alpha)$, since $K \succeq \hat{K}(K,\mathcal{C})$. To the best of our knowledge, only a weaker result about the accuracy of the Nyström approximation [16] related to k-DPPs exists in the literature.

2.2. Unsupervised kernel methods . Theorem 2.4 states that the 'distance' of K to a k-dimensional subspace is well approximated by the 'distance' of $\hat{K}(K,\mathcal{C})$ to the same subspace, on expectation. An analogous results of RLS sampling can be found in [18].

THEOREM 2.4 (Expected projection-cost preservation). *Let $\mathcal{C} \sim DPP_L(K/\alpha)$ and X an orthogonal projector on a k-dimensional subspace. Denote $\hat{K} = \hat{K}(K,\mathcal{C})$. Then we have*

$$\mathrm{Tr}(K - XKX) \leq \mathbb{E}_{\mathcal{C}}[\mathrm{Tr}(\hat{K} - X\hat{K}X)] + c(\alpha) \leq \mathrm{Tr}(K - XKX) + \min\{\alpha k, c(\alpha)\},$$

where $c(\alpha) = \alpha d_{\mathrm{eff}}(K/\alpha)$.

A direct application of the above theorem is KPCA. Namely, the projector $X^\star$ onto the leading k components is obtained by solving

$$\min_{X \in \Pi_k} \mathrm{Tr}(K - XKX),$$

where Π_k is the set of $n \times n$ projectors of rank k. Then, the result Theorem 2.4 is a stability result relating the objective functions of KPCA with and without Nyström approximation.

2.3. Kernel Ridge Regression . A simple consequence of Corollary 2.3 is that the expected risk of KRR approximated by Nyström method with DPP sampling cannot be arbitrary larger than the risk corresponding to the full KRR. Namely, let the observed outputs be $y_i = z_i + \epsilon_i$ where ϵ_i are iid $\mathcal{N}(0,\sigma^2)$ and z_i the noiseless output. Let the solution of KRR be $\hat{\boldsymbol{z}}_K = K(K + n\gamma\mathbb{I})^{-1}\boldsymbol{y}$, and denote by $\hat{\boldsymbol{z}}_{\hat{K}(K,\mathcal{C})}$ the solution obtained by using the Nyström approximation $\hat{K}(K,\mathcal{C})$. The expected risk is then defined as $\mathcal{R}(\hat{\boldsymbol{z}}_K) = \mathbb{E}_\epsilon\|\hat{\boldsymbol{z}}_K - \boldsymbol{z}\|_2^2$. Then, we can give a bound on the risk of KRR associated to the Nyström approximation.

THEOREM 2.5 (Expected risk bound). *Let $\mathcal{C} \sim DPP_L(K/\alpha)$, then we have*

$$\mathbb{E}_{\mathcal{C}}\left[\sqrt{\frac{\mathcal{R}(\hat{\boldsymbol{z}}_{\hat{K}(K,\mathcal{C})})}{\mathcal{R}(\hat{\boldsymbol{z}}_K)}}\right] \leq 1 + \frac{\alpha}{n\gamma}d_{\mathrm{eff}}(K/\alpha).$$

The upper bound in Theorem 2.5 tends to 1 as $\alpha \to 0$ since $d_{\text{eff}}(K/\alpha) \leq n$. This consistently shows that the larger is the number of landmarks, the closest is the risk of approximate KRR from the full KRR. Notice that the increase in the risk is also mitigated by the regularization parameter.

Preconditioners. Rudi et al. [23] propose a preconditioning of the linear system (1.5) of the form

$$B^\top (K_{\mathcal{C}}^\top K_{\mathcal{C}} + n\gamma K_{\mathcal{C}\mathcal{C}}) B\big(B^{-1}\boldsymbol{\alpha}\big) = B^\top K_{\mathcal{C}}^\top \boldsymbol{y} \tag{2.2}$$

where B is obtained by solving $BB^\top = \big(K_{\mathcal{C}\mathcal{C}} D_{\mathcal{C}\mathcal{C}} K_{\mathcal{C}\mathcal{C}} + n\gamma K_{\mathcal{C}\mathcal{C}}\big)^{-1}$, thanks to a Cholesky decomposition, where $D_{\mathcal{C}\mathcal{C}}$ is an appropriate diagonal matrix. In the case of the uniform sampling of $\mathcal{C}$, the authors of [23] propose $D_{\mathcal{C}\mathcal{C}} = (n/|\mathcal{C}|)\mathbb{I}_{\mathcal{C}\mathcal{C}}$. For RLS sampling, they argue for $D_{\mathcal{C}\mathcal{C}} = \text{Diag}(\boldsymbol{\ell}_{\mathcal{C}})^{-1}$, where $\boldsymbol{\ell}$ contains the so-called ridge leverage scores. We emphasize that the computation of B indeed crucially depends on the magnitude of the least eigenvalue of $K_{\mathcal{C}\mathcal{C}}$. It is then interesting to sample diverse landmarks so that $K_{\mathcal{C}\mathcal{C}}$ is likely to be regular. A natural motivation for choosing $D_{\mathcal{C}\mathcal{C}} = \text{Diag}(\boldsymbol{\ell}_{\mathcal{C}})^{-1}$ from the DPP viewpoint is given in Corollary 2.7. This result naturally follows from Lemma 2.6, which can also be found in the context of Monte-Carlo integration [3] with projective DPPs.

LEMMA 2.6. *Let $\mathcal{C} \sim DPP_L(K/\alpha)$ and $\boldsymbol{v}$ and $\boldsymbol{w} \in \mathbb{R}^n$ Then, we have the identities $\mathbb{E}_{\mathcal{C}}\left[\boldsymbol{v}_{\mathcal{C}}^\top \boldsymbol{w}_{\mathcal{C}}\right] = \boldsymbol{v}^\top \text{Diag}(\boldsymbol{\ell})\boldsymbol{w}$, and $\mathbb{V}_{\mathcal{C}}\left[\boldsymbol{v}_{\mathcal{C}}^\top \boldsymbol{w}_{\mathcal{C}}\right] = (\boldsymbol{v}\circ\boldsymbol{w})^\top \Big(\text{Diag}(\boldsymbol{\ell}) - P\circ P\Big)(\boldsymbol{v}\circ\boldsymbol{w})$.*

Corollary 2.7 then motivates the approximation of $K_{\mathcal{C}}^\top K_{\mathcal{C}}$ in (1.5) by $K_{\mathcal{C}\mathcal{C}} D_{\mathcal{C}\mathcal{C}} K_{\mathcal{C}\mathcal{C}}$.

COROLLARY 2.7. *Let $\mathcal{C} \sim DPP_L(K/\alpha)$ and $K_{\mathcal{C}} = KC$. Then, the following identity holds: $\mathbb{E}_{\mathcal{C}}\left[K_{\mathcal{C}} \,\text{Diag}(\boldsymbol{\ell}_{\mathcal{C}})^{-1} K_{\mathcal{C}}^\top\right] = K^2$.*

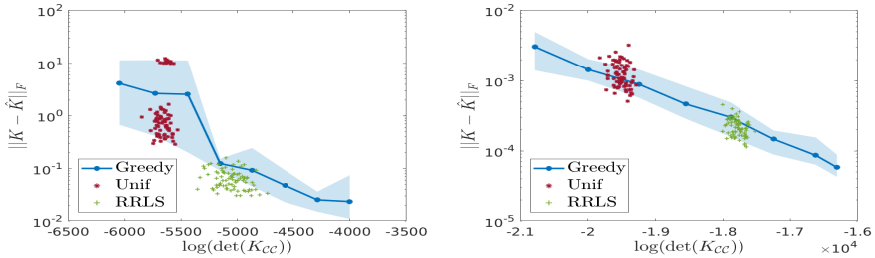
A formula for the variance can also be obtained thanks to Lemma 2.6. Again, using DPP sampling with inverse leverage score preconditioning has the advantage that $K_{\mathcal{C}\mathcal{C}}$ is likely to be regular, in contrast with leverage score sampling.

3. Experimental results. In this section, we illustrate the effect of sampling a subset $\mathcal{C}$ with small or large $\det(K_{\mathcal{C}\mathcal{C}})$ on a number of public datasets. A swapping algorithm, described in Algorithm 2.1, is used to determine subsets of size $|\mathcal{C}| = k$ with a given log-determinant d_p, such that $|\log(\det(K_{\mathcal{C}\mathcal{C}})) - d_p| \leq \epsilon$ and where $\epsilon > 0$ is a numerical tolerance. The method swaps points in and out of an initial subset, so that the swapped point is accepted if the determinant of the new submatrix is closer to the desired determinant d_p. If the determinant of the subset is too small, we sample a new candidate by using (approximate) leverage scores sampling. Otherwise, if the determinant is too large, we use inverse leverage scores sampling. The size of the subset is chosen to be the effective dimension $k = \sum_{i=1}^n \boldsymbol{\ell}_i$, where $\{\boldsymbol{\ell}_i\}_{i=1}^n$ are the ridge leverage scores with regularization parameter $\alpha = \lambda n$. The same ridge leverage scores are used in the greedy swapping algorithm. The algorithm stops if the desired precision is reached or the number of iterations exceeds 2000, whichever happens earlier. For large-scale problems, the ridge leverage scores are approximated using Recursive Ridge Leverage Sampling (RRLS) [18] with n_{RRLS} points. The size of the subset is chosen to be the effective dimension of the approximate RLS with regularization parameter $\alpha = \lambda n$. The same approximate RLS are used in the greedy swapping algorithm, where the maximum number of iterations is set to 5000. It is worth mentioning that a faster approximation of RLS can be obtained by BLESS [22], although it requires a bit more parameter tuning compared with RRLS. This is why the latter method has been selected for RLS sampling in this paper.

Algorithm 2.1 Greedy Swapping Algorithm based on the (approximated) ridge leverage scores.

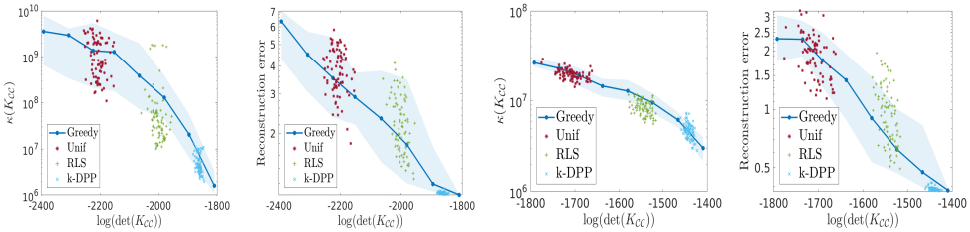
input: Matrix $K \succ 0$, sample size k, ridge leverage scores $\{\boldsymbol{\ell}_i\}_{i=1}^n$, preferred log-determinant d_p and precision $\epsilon > 0$.
initialization: Sample an initial subset $|\mathcal{C}| = k$ uniformly at random.
Determine the Cholesky decomposition R, with $K_{\mathcal{C}\mathcal{C}} = R^\top R$.
repeat:
  Determine the log-determinant $d = 2\sum_{i=1}^k \log(R_{ii})$.
  if: $|d - d_p| \leq \epsilon$
    break
  if: $d < d_p$
    Sample a new point $\tilde{c}$ out of the remaining subset with $p_i \sim \boldsymbol{\ell}_i$
  otherwise:
    Sample a new point $\tilde{c}$ out of the remaining subset with $p_i \sim 1 - \boldsymbol{\ell}_i$
  Swap a uniform selected point out of $\mathcal{C}$ with the newly sampled $\tilde{c}$, which gives the new subset $\tilde{\mathcal{C}}$
  Do a rank-1 update to the Cholesky decomposition, which gives $\tilde{R}$ and determine $\tilde{d} = 2\sum_{i=1}^k \log(\tilde{R}_{ii})$.
  if: $|\tilde{d} - d_p| \leq |d - d_p|$
    Keep the swapped point and update $R = \tilde{R}$, $\mathcal{C} = \tilde{\mathcal{C}}$
return $\mathcal{C}$.

Settings. In the sequel, a Gaussian kernel with bandwidth σ is used after standardizing the data. All the simulations are repeated 10 times, the averaged is displayed and the errorbars show the 0.05 and 0.95 quantile. The datasets and hyperparameters are given in Table 1. In the first case-studies, the following exact algorithms are used to sample k landmarks: Uniform sampling (Unif.), Ridge Leverage Score sampling (RLS) [9] and k-DPP [14]. For a fair comparison, we chose to use k-DPP rather than DPP so that the number of landmarks is always constant and equal to the expected subset size of the associated DPP. In the large-scale experiments: Unif., RRLS and the greedy swapping method are compared.



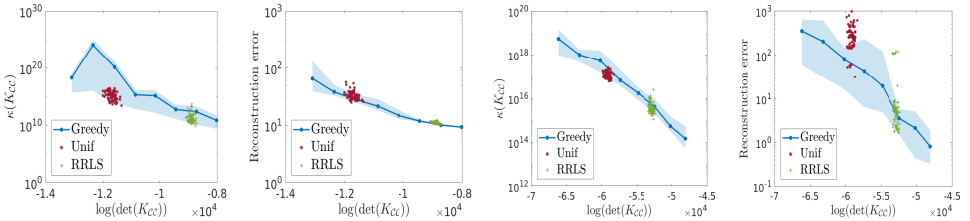
(a) `MiniBooNE`: error (b) `codRNA`: error

Fig. 3: Large-scale Nyström approximation results. The Frobenius norm of the approximation error of $\hat{K} = \hat{K}(K, \mathcal{C})$ is plotted versus the sample diversity. The larger $\det(K_{\mathcal{C}\mathcal{C}})$, the more diverse the subset.



(a) `A.Credit`: $\kappa(K_{\mathcal{CC}})$ (b) `A.Credit`: error (c) `B.Cancer`: $\kappa(K_{\mathcal{CC}})$ (d) `B.Cancer`: error

Fig. 4: KPCA results. The condition number, minimum eigenvalues and reconstruction error using half of the components are plotted as a function of $\det(K_{\mathcal{CC}})$.



(a) `Adult`: $\kappa(K_{\mathcal{CC}})$ (b) `Adult`: error (c) `Cov`: $\kappa(K_{\mathcal{CC}})$ (d) `Cov`: error

Fig. 5: Large-scale KPCA results. The condition number and reconstruction error using half of the components are plotted as a function of $\det(K_{\mathcal{CC}})$.

Nyström approximation. The impact of diversity on the Nyström approximation is illustrated on the `Housing`, `Abalone`, `codRNA` and `MiniBooNE` datasets[2]. The condition number of $K_{\mathcal{CC}}$, its largest/smallest eigenvalues and the accuracy of the Nyström approximation are plotted as a function of the determinant in Figures 1 and 3.

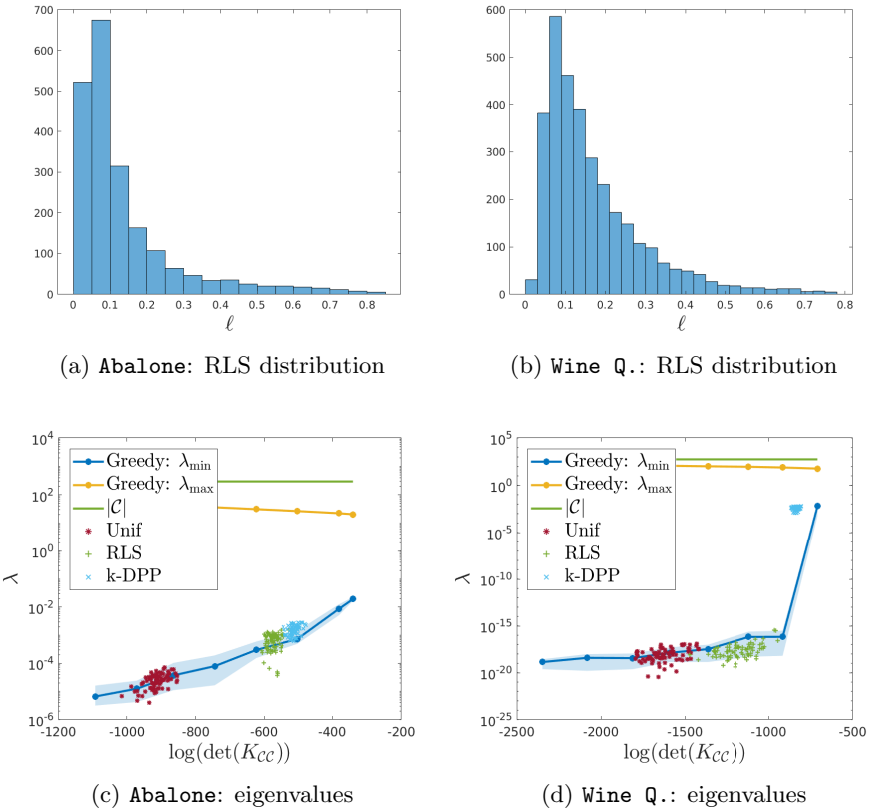
For completeness, the largest and smallest eigenvalues of $K_{\mathcal{CC}}$ are also given in the appendix. The accuracy of the approximation is evaluated by calculating $\|K - \hat{K}\|_F / \|K\|_F$ with $\hat{K} = KC(K_{\mathcal{CC}} + \varepsilon \mathbb{I}_{\mathcal{CC}})^{-1} C^\top K$ with $\varepsilon = 10^{-12}$ for numerical stability.

Afterwards, the following algorithms are used to sample k landmarks: Uniform sampling (Unif.), exact ridge leverage score sampling (RLS) [9] and k-DPP [14]. The results in Figure 1 show that the 3 sampling algorithms follow the general trend of the greedy swapping algorithm, namely, we have to following empirical observations: **1)** Sampling a more diverse subset results in a smaller condition number $\kappa(K_{\mathcal{CC}})$. This is mainly because a larger determinant corresponds to a larger $\lambda_{\min}(K_{\mathcal{CC}})$. Indeed, diverse sampling is a computational regularization. **2)** Sampling a diverse subset gives a more accurate Nyström approximation. In practice, we observe that RLS sampling yields *effectively* more diverse samples compared to uniform sampling. Notice that in the presence of outliers, taking samples with an extremely large $\det(K_{\mathcal{CC}})$ thanks to the Greedy Swapping Algorithm might increase the error on the Nyström approximation

[2] https://www.cs.toronto.edu/~delve/data/datasets.html, https://archive.ics.uci.edu/ml/index.php

as it explained in Supplementary Material. The results for the large-scale experiments are visualized on Figure 3. The accuracy of the approximation is now evaluated by averaging the Frobenius norm error $\|K - \hat{K}\|_F$ over 50 subsets of size 3000.

Kernel PCA. The numerical experiments are done on the `Breast Cancer` (`B. Cancer`), `Australian Credit` (`A. Credit`), `Adult` and `Covertype` datasets[1]. The condition number, smallest/largest eigenvalues of $K_{\mathcal{CC}}$ and the reconstruction error are plotted as a function of the determinant. The averaged results are visualized in Figure 4. Information about the datasets and hyperparameters used for the experiments is given in Table 1 in appendix. Empirically, sampling a more diverse subset results in a smaller reconstruction error and a smaller condition number. The results for the large-scale experiments are visualized on Figure 5.



(a) `Abalone`: RLS distribution

(b) `Wine Q.`: RLS distribution

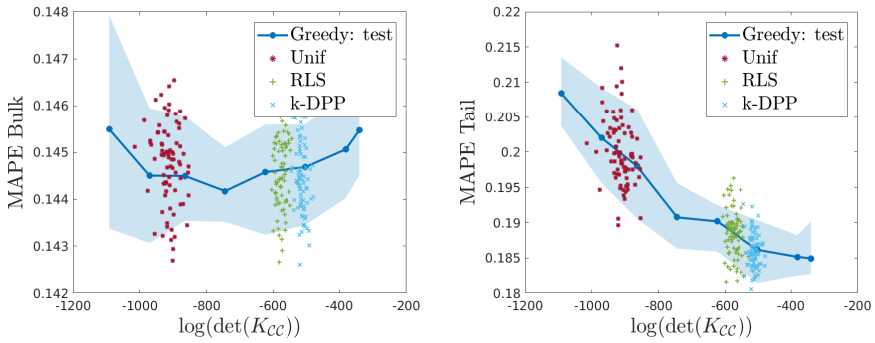
(c) `Abalone`: eigenvalues

(d) `Wine Q.`: eigenvalues

Fig. 6: Regression results. The histogram of the RLS of the test set and minimum and maximum eigenvalues of $K_{\mathcal{CC}}$ versus $\det(K_{\mathcal{CC}})$, accompanying Figure 7.

Regression and Stratification of the error. To conclude, we verify the usefulness of diversity for a supervised learning task. The dataset is split in 50% training data and 50% test data, so to make sure the train and test set have similar RLS distributions. The test RLS distribution is visualized in Figure 6. The regression experiment is repeated on the `Abalone`, `Wine Quality`, `Bike Sharing` (`Bike S.`) and `YearPredictionMSD` (`Year`) datasets[1] by using KRR. The MAPE of the kernel ridge

regression is calculated as a function of $\det(K_{\mathcal{CC}})$. To evaluate the performance, the dataset is stratified, i.e., the test set is divided into 'bulk' and 'tail' as follows: the bulk corresponds to test points where the RLS are smaller than or equal to the 70% quantile, while the tail of the data corresponds to test points where the ridge leverage score is larger than the 70% quantile. The regularization parameter γ is determined by using cross-validation. The results in Figure 7 show that the 3 sampling algorithms follow the general trend of the greedy swapping algorithm. Again here, sampling a more diverse subset, results in a better conditioning of the kernel sub-matrix. Diverse sampling has comparable performance for the bulk data, while performing much better in the tail of the data. This confirms the expectations from the regression toy example. The results for the large-scale experiments are visualized on Figure 8, where the Symmetric MAPE (SMAPE) is shown in the bulk and tail of the data.

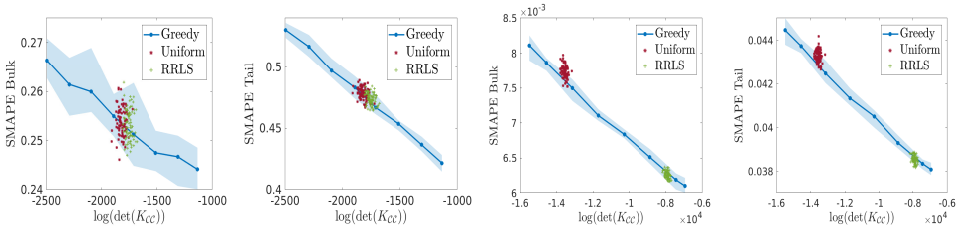


(a) `Abalone`: MAPE Bulk (b) `Abalone`: MAPE Tail

(c) `Wine Q.`: MAPE Bulk (d) `Wine Q.`: MAPE Tail

Fig. 7: Regression results. The MAPE on the test set is plotted as a function of $\det(K_{\mathcal{CC}})$. A small MAPE corresponds to a large accuracy.

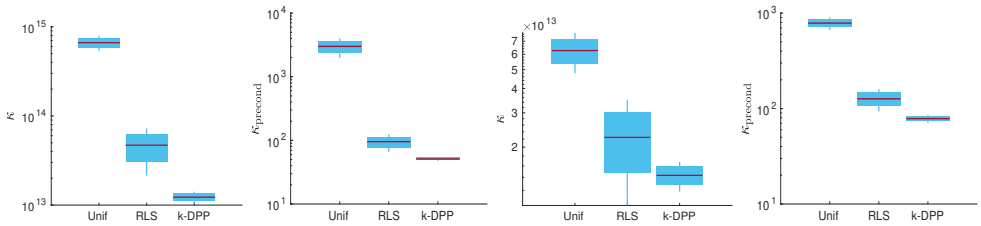
Preconditioner. A natural idea is to consider iterative methods to solve the system in (1.5) because of their simplicity and low iteration cost. The speed and accuracy of convergence of the conjugate gradient method depends on the condition number of the linear system [13], which makes the use of diverse samplings in combination with iterative methods particularly interesting. This is illustrated on the `Parkinson`, and



(a) `Bike S.`: Bulk (b) `Bike S.`: Tail (c) `Year Pred.`: Bulk (d) `Year Pred.`: Tail

Fig. 8: Large-scale Regression results. The SMAPE on the bulk and tail of the test set versus the sample diversity. The larger $\det(K_{\mathcal{C}\mathcal{C}})$, the more diverse the subset.

`Pumadyn8FM` datasets[1]. The condition number of the preconditioned system is com-



(a) `Parkinson` (b) `Parkinson` (c) `Pumadyn8FM` (d) `Pumadyn8FM`

Fig. 9: Preconditioning results. The condition number of the linear system before and after the preconditioning is plotted for Uniform, RLS and k-DPP sampling. From left to right, the condition number before and after preconditioning, for `Parkinson` and `Pumadyn8FM` datasets, respectively.

puted, where the preconditionner given in (2.2) is used in combination with uniform sampling, RLS and k-DPP in combination with the preconditioner defined in (2.2). The ridge regularization parameter is equal to $\lambda = 10^{-10}$ to illustrate the impact of diversity on badly conditioned systems. From the results in Figure 9, we see empirically that sampling a more diverse subset, results in a better conditioning of the linear system.

4. Conclusions. In this paper, the interest of sampling diverse landmarks in the context of Nyström approximation was illustrated. Our empirical findings relating regularization and diversity are partly supported by theoretical results. In the context of Kernel Ridge Regression, an extra contribution of the paper consists in proposing the use of the Ridge Leverage Score distribution in order to assess the uniformity of the performance of the regressor.

Appendix A. Proofs.

Proof of Theorem 2.1. Without loss of generality, we put $\alpha = 1$.
(i) Then, we first prove $\mathbb{E}_{\mathcal{C}}[CK_{\mathcal{C}\mathcal{C}}^{-1}C^{\top}] = (K + \mathbb{I})^{-1}$. The matrix inversion lemma

yields

$$\det\begin{pmatrix} K_{\mathcal{C}\mathcal{C}} & C^\top \boldsymbol{w} \\ \boldsymbol{u}^\top C & 1 \end{pmatrix} = \det(K_{\mathcal{C}\mathcal{C}}) \left[1 - \boldsymbol{u}^\top C K_{\mathcal{C}\mathcal{C}}^{-1} C^\top \boldsymbol{w}\right] = \det\left(K_{\mathcal{C}\mathcal{C}} - C^\top \boldsymbol{w}\boldsymbol{u}^\top C\right).$$

This simplifies to

$$\boldsymbol{u}_{\mathcal{C}}^\top K_{\mathcal{C}\mathcal{C}}^{-1} \boldsymbol{w}_{\mathcal{C}} = 1 - \frac{\det(K_{\mathcal{C}\mathcal{C}} - C^\top \boldsymbol{w}\boldsymbol{u}^\top C)}{\det(K_{\mathcal{C}\mathcal{C}})}.$$

By taking the expectation on both sides, we find

$$\mathbb{E}_{\mathcal{C}}\left[\boldsymbol{u}_{\mathcal{C}}^\top K_{\mathcal{C}\mathcal{C}}^{-1} \boldsymbol{w}_{\mathcal{C}}\right] = 1 - \frac{\det(\mathbb{I} + K - \boldsymbol{w}\boldsymbol{u}^\top)}{\det(\mathbb{I} + K)} = \boldsymbol{u}^\top (\mathbb{I} + K)^{-1} \boldsymbol{w},$$

where we used $\sum_{\mathcal{C}\subseteq[n]} A_{\mathcal{C}\mathcal{C}} = \mathbb{I} + A$ with A a square matrix.
(ii) Secondly, we prove that $\mathbb{E}_{\mathcal{C}}[\boldsymbol{u}^\top CC^\top KCC^\top \boldsymbol{v}] = P_{(2)} \circ K$. We first calculate by using the marginal kernel

$$\mathbb{E}_{\mathcal{C}}[1_{i\in\mathcal{C}} 1_{j\in\mathcal{C}}] = 1_{i=j} \times P_{ii} + 1_{i\neq j} \times \det P_{\{i,j\}\{i,j\}} = 1_{i=j} \times P_{ii} + (P_{ii}P_{jj} - P_{ij}^2).$$

The result follows from the following formula

$$\mathbb{E}_{\mathcal{C}}[\boldsymbol{u}^\top CC^\top KCC^\top \boldsymbol{v}] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}_{\mathcal{C}}[u_i 1_{i\in\mathcal{C}} K_{ij} 1_{j\in\mathcal{C}} v_j],$$

and by using the linearity of the expectation. ∎

Proof of Corollary 2.2. We follow the proof strategy of [16]. Let the following $f(\mathcal{C}) = \boldsymbol{w}_{\mathcal{C}}^\top K_{\mathcal{C}\mathcal{C}}^{-1} \boldsymbol{w}_{\mathcal{C}}$ a function viewed as $f : \{0,1\}^n \to \mathbb{R}$. Firstly, we use a concentration inequality, then we upper bound the Lipschitz constant of f.
Concentration inequality. To start with, let us denote by N the number of points sampled by a DPP. Then, we quote a concentration result from [18].

THEOREM A.1 (Thm 3.5 in [21]). *Let a determinantal point process with $\mathbb{E}N = \mu < \infty$. Let f be a 1-Lipschitz function on finite counting measures. Then ,*

$$\Pr(|f - \mathbb{E}f| > a) \leq 5\exp\left(-\frac{a^2}{16(a+2\mu)}\right).$$

Let $\delta > 0$ be an upper bound on the failure probability, such that $5\exp\left(-\frac{a^2}{16(a+2\mu)}\right) \geq \delta$ for some $a > 0$. This yields to the inequality

$$a^2 - 16a\log\left(\frac{5}{\delta}\right) - 32\mu\log\left(\frac{5}{\delta}\right) \geq 0.$$

Therefore, we find the lower bound

$$a \geq 8\left(\log\left(\frac{5}{\delta}\right) + \sqrt{\left(\log\left(\frac{5}{\delta}\right)\right)^2 + \frac{\mu}{2}\log\left(\frac{5}{\delta}\right)}\right). \tag{A.1}$$

Then, notice that we can use the above Theorem for ℓ-Lipshitz function by simply replacing f by f/ℓ. Hence, with a probability at least δ, we have

$$\begin{aligned}|f - \mathbb{E}f| \leq & \ell 8 \left(\log\left(\frac{5}{\delta}\right) + \sqrt{\left(\log\left(\frac{5}{\delta}\right)\right)^2 + \frac{\mu}{2}\log\left(\frac{5}{\delta}\right)} \right) \\ & \leq 8\ell \left(2\log\left(\frac{5}{\delta}\right) + \sqrt{\frac{\mu}{2}\log\left(\frac{5}{\delta}\right)} \right) \\ & \leq 8\ell \left(2 + \sqrt{\frac{\mu}{2}} \right) \log\left(\frac{5}{\delta}\right),\end{aligned}$$

where we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for positive a and b, and that $\log\left(\frac{5}{\delta}\right) \geq 1$. *Bounding the Lipschitz constant.* We now calculate an upper bound on the Lipschitz constant of f with respect to the Hamming distance. To do so, it is sufficient to consider two binary vectors C and $C' \in \{0,1\}^n$ differing of one digit. Say that $f(C) \geq f(C') \geq 0$. Then, we have

$$|f(C) - f(C')| \leq f(C) - f(C') \leq \boldsymbol{w}_{\mathcal{C}}^\top K_{\mathcal{C}\mathcal{C}}^{-1} \boldsymbol{w}_{\mathcal{C}} \leq \frac{1}{\lambda_{\min}(K)} = \frac{d_H(C,C')}{\lambda_{\min}(K)},$$

since $\|\boldsymbol{w}\|_2 = 1$ and where $d_H(C,C')$ is the Hamming distance between the binary vectors C and C'. Notice that the inequality above used the following fact

$$K_{\mathcal{C}\mathcal{C}}^{-1} \preceq \frac{1}{\lambda_{\min}(K_{\mathcal{C}\mathcal{C}})} \mathbb{I}_{\mathcal{C}\mathcal{C}},$$

and $\lambda_{\min}(K) \leq \lambda_{\min}(K_{\mathcal{C}\mathcal{C}})$ (see [12]), as well as $CC^\top \preceq \mathbb{I}$. The final result follows by taking $\ell = 1/\lambda_{\min}(K)$. ∎

Proof of Theorem 2.4. Since X is a orthogonal projector and $L_{\mathcal{C}} \preceq K$, we have $\operatorname{Tr}(XL_{\mathcal{C}}X) \leq \operatorname{Tr}(XKX)$, which yields

$$\operatorname{Tr}(K - XKX) \leq \operatorname{Tr}(L_{\mathcal{C}} - XL_{\mathcal{C}}X) + \operatorname{Tr}(K - L_{\mathcal{C}}) = \operatorname{Tr}(K) - \operatorname{Tr}(XL_{\mathcal{C}}X).$$

By taking the expectation over $\mathcal{C} \sim DPP_L(K/\alpha)$ on both sides of the above inequality, and by using $\mathbb{E}_{\mathcal{C}}(L_{\mathcal{C}}) = K - \alpha P$, we obtain

$$\operatorname{Tr}(K - XKX) \leq \mathbb{E}_{\mathcal{C}}[\operatorname{Tr}(L_{\mathcal{C}} - XL_{\mathcal{C}}X)] + \alpha d_{\text{eff}}(K/\alpha) = \operatorname{Tr}(K - XKX) + \alpha \operatorname{Tr}(XPX).$$

Finally, by using that X is a projector, it holds that $\operatorname{Tr}(XPX) \leq \operatorname{Tr}(X) = k$ since $P \preceq \mathbb{I}$. Also, we have $\operatorname{Tr}(XPX) \leq \operatorname{Tr}(P) = d_{\text{eff}}(K/\alpha)$, so that the final bound is obtained. ∎

Proof of Theorem 2.5. The risk $\mathcal{R}(\hat{z}) = \frac{1}{n}\mathbb{E}_\epsilon \|\hat{z} - z\|_2^2$ is decomposed in terms of bias and variance $\mathcal{R}(\hat{z}) = \text{bias}^2(K) + \text{var}(K)$, where the bias reads

$$\text{bias}(K) = \sqrt{n\gamma^2 z^\top (K + n\gamma\mathbb{I})^{-2} z}$$

and the variance is

$$\text{var}(K) = \frac{\sigma^2}{n} \operatorname{Tr}[K^2 (K + n\gamma\mathbb{I})^{-2}].$$

Firstly, since $L_{\mathcal{C}} \preceq K$, it holds that $\lambda_\ell(L_{\mathcal{C}}) \leq \lambda_\ell(K)$ for all ℓ and $\mathrm{var}(L_{\mathcal{C}}) \preceq \mathrm{var}(K)$. Then, we can give an upper bound for $\mathrm{bias}(L_{\mathcal{C}})$. The 2-norm submultiplicativity gives

$$\begin{aligned}\|(L_{\mathcal{C}} + n\gamma\mathbb{I})^{-1}z - (K + n\gamma\mathbb{I})^{-1}z\|_2 &\leq \|(L_{\mathcal{C}} + n\gamma\mathbb{I})^{-1}(K - L_{\mathcal{C}})\|_2 \|(K + n\gamma\mathbb{I})^{-1}z\|_2 \\ &\leq \frac{\mathrm{Tr}(K - L_{\mathcal{C}})}{n\gamma}\|(K + n\gamma\mathbb{I})^{-1}z\|_2.\end{aligned}$$

Hence, by using the triangle inequality and the bound hereabove, it holds that

$$\begin{aligned}\sqrt{z^\top(L_{\mathcal{C}} + n\gamma\mathbb{I})^{-2}z} &\leq \|(K + n\gamma\mathbb{I})^{-1}z\|_2 + \|(L_{\mathcal{C}} + n\gamma\mathbb{I})^{-1}z - (K + n\gamma\mathbb{I})^{-1}z\|_2 \\ &\leq (1 + \frac{\mathrm{Tr}(K - L_{\mathcal{C}})}{n\gamma})\sqrt{z^\top(K + n\gamma\mathbb{I})^{-2}z}.\end{aligned}$$

The result follows by taking the expectation over $\mathcal{C} \sim DPP_L(K/\alpha)$. ∎

Proof of Lemma 2.6. This results is a direct consequence of the two following identities: $\mathbb{E}_{\mathcal{C}}[1_{i\in\mathcal{C}}] = P_{ii}$, and

$$\mathbb{E}_{\mathcal{C}}[1_{i\in\mathcal{C}}1_{j\in\mathcal{C}}] = 1_{i=j} \times P_{ii} + (P_{ii}P_{jj} - P_{ij}^2).$$ ∎

Appendix B. Parameters and dataset descriptions. The parameters and datasets used in the simulations can be found in Table 1. When a subset is sampled from k-DPP, the number of landmarks is fixed to $k = d_{\mathrm{eff}}(n\lambda)$, where the effective dimension corresponds to the expected subset size for $DPP_L(K/(n\lambda))$.

Table 1: Datasets and parameters used for the experiments on the Nyström approximation.

Dataset	Task	n	d	σ	λ	$d_{\mathrm{eff}}(n\lambda)$	n_{RRLS}
`Housing`	Kernel approx.	506	13	5	10^{-6}	186	/
`MiniBooNE`	Kernel approx.	130065	50	8	10^{-6}	462	8000
`codRNA`	Kernel approx.	331152	8	8	10^{-6}	1204	8000
`B. Cancer`	KPCA	569	30	10	10^{-6}	158	/
`A. Credit`	KPCA	690	14	5	10^{-6}	371	/
`Adult`	KPCA	48842	110	8	10^{-6}	1202	8000
`Covertype`	KPCA	581012	54	20	10^{-6}	3665	10000
`Abalone`	KRR	4177	8	1	10^{-4}	294	/
`Wine Quality`	KRR	6497	11	2	10^{-4}	555	/
`Bike S.`	KRR	17389	16	1	10^{-4}	2732	8000
`Year Pred.`	KRR	515345	90	10	10^{-4}	4217	10000
`Parkinson`	Prec.	5875	20	5	10^{-6}	738	/
`Pumadyn8FM`	Prec.	8192	25	5	10^{-6}	296	/

Acknowledgments. We thank Michał Dereziński and Michael Mahoney for pointing out the reference [19].

REFERENCES

[1] F. Bach, *Sharp analysis of low-rank kernel matrix approximations*, in COLT Conference on Learning Theory, 2013, pp. 185–209.
[2] R. Bapat and V. Sunder, *On majorization and Schur products*, Linear Algebra and its Applications, 72 (1985), pp. 107 – 117.
[3] R. Bardenet and A. Hardy, *Monte Carlo with Determinantal Point Processes*, Annals of Applied Probability, (2019).
[4] M. Belkin, D. J. Hsu, and P. Mitra, *Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate*, in Advances in Neural Information Processing Systems, 2018, pp. 2300–2311.
[5] M. Belkin, S. Ma, and S. Mandal, *To understand deep learning we need to understand kernel learning*, arXiv preprint arXiv:1802.01396, (2018).
[6] V. Chen, S. Wu, A. J. Ratner, J. Weng, and C. Ré, *Slice-based learning: A programming model for residual learning in critical data slices*, in Advances in neural information processing systems, 2019, pp. 9392–9402.
[7] A. Civril and M. Magdon-Ismail, *On selecting a maximum volume sub-matrix of a matrix and related problems*, Theoretical Computer Science, 410 (2009), pp. 4801 – 4811.
[8] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, *Matrix approximation and projective clustering via volume sampling*, in Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, Society for Industrial and Applied Mathematics, 2006, pp. 1117–1126.
[9] A. El Alaoui and M. Mahoney, *Fast randomized kernel ridge regression with statistical guarantees*, in Advances in Neural Information Processing Systems 28, 2015, pp. 775–783.
[10] M. Fanuel, J. Schreurs, and J. A. Suykens, *Nyström landmark sampling and regularized Christoffel functions*, arXiv preprint arXiv:1905.12346, (2019).
[11] V. Feldman, *Does learning require memorization? a short tale about a long tail*, arXiv preprint arXiv:1906.05271, (2019).
[12] C. Johnson and H. Robinson, *Eigenvalue inequalities for principal submatrices*, Linear Algebra and its Applications, 37 (1981), pp. 11 – 22.
[13] D. S. Kershaw, *The incomplete Cholesky—conjugate gradient method for the iterative solution of systems of linear equations*, Journal of computational physics, 26 (1978), pp. 43–65.
[14] A. Kulesza and B. Taskar, *k-DPPs: Fixed-size determinantal point processes*, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 1193–1200.
[15] A. Kulesza and B. Taskar, *Determinantal point processes for machine learning*, Foundations and Trends in Machine Learning, 5 (2012), pp. 123–286.
[16] C. Li, S. Jegelka, and S. Sra, *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, 2016, pp. 2061–2070.
[17] T. Liang and A. Rakhlin, *Just interpolate: Kernel" ridgeless" regression can generalize*, arXiv preprint arXiv:1808.00387, (2018).
[18] C. Musco and C. Musco, *Recursive Sampling for the Nyström Method*, in Advances in Neural Information Processing Systems 30, 2017, pp. 3833–3845.
[19] M. Mutny, M. Derezinski, and A. Krause, *Convergence analysis of block coordinate algorithms with determinantal sampling*, in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 108, PMLR, 2020, pp. 3110–3120.
[20] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging*, arXiv preprint arXiv:1909.12475, (2019).
[21] R. Pemantle and Y. Peres, *Concentration of lipschitz functionals of determinantal and other strong rayleigh measures*, Combinatorics, Probability and Computing, 23 (2014), pp. 140–160.
[22] A. Rudi, D. Calandriello, L. Carratino, and L. Rosasco, *On fast leverage score sampling and optimal learning*, in Advances in Neural Information Processing Systems, 2018, pp. 5673–5683.
[23] A. Rudi, L. Carratino, and L. Rosasco, *Falkon: An optimal large scale kernel method*, in Advances in Neural Information Processing Systems 30, 2017, pp. 3888–3898.
[24] B. Schölkopf, A. Smola, and K.-R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural computation, 10 (1998), pp. 1299–1319.
[25] N. Sterge, B. Sriperumbudur, L. Rosasco, and A. Rudi, *Gain with no Pain: Efficient Kernel-PCA by Nyström Sampling*, arXiv preprint arXiv:1907.05226, (2019).
[26] F. J. Valverde-Albacete and C. Peláez-Moreno, *100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox*, PloS

one, 9 (2014), p. e84217.
[27] S. Wang, A. Gittens, and M. W. Mahoney, *Scalable kernel K-means clustering with Nyström approximation: relative-error bounds*, The Journal of Machine Learning Research, 20 (2019), pp. 431–479.

SUPPLEMENTARY MATERIALS: DIVERSITY SAMPLING IS AN IMPLICIT REGULARIZATION FOR KERNEL METHODS *

MICHAËL FANUEL†, JOACHIM SCHREURS†, AND JOHAN A.K. SUYKENS†

Setting. In the performance plots displayed in the sequel, i.e., Figures SM2, SM3, SM4, SM5, SM6, SM7 and SM8, the results are plotted on a logarithmic scale, averaged over 10 trials and the errobars show the 0.05 and 0.95 quantile. Recall that the larger the $\log(\det(K_{CC}))$, the more diverse the subset. The computer used for the small-scale simulations has 8 cores 3.40GHz and 15.5 GB of RAM. Large scale experiments on `Covertype` and `Year PredictionMSD` were done on the Vlaams Super Computer (VSC). The implementation of the algorithms is done with matlabR2018b.

SM0.1. Additional case study: Clustering. The performance of diverse kernel approximation methods is evaluated for a clustering task using kernel k-means with Nyström approximation [SM2]. Samples with different diversities are sampled, afterwards *Algorithm 2* of Wang et al. [SM2] is used to cluster the dataset, where the target dimension $s = k$ is always equal to the number of desired clusters. The clustering performance is evaluated by the normalized mutual information (NMI) [SM1], the NMI gives a value between 0 and 1, where 1 represents perfect correlation between the ground truth and the clustering outcome. We first illustrate the effect of using a diverse sampling on a toy example. In Figure SM1, we show a very imbalanced

Table SM1: Datasets and parameters used in the clustering experiments.

Dataset	n	d	σ	λ	$d_{\text{eff}}(n\lambda)$	# Clusters
`Glass`	214	9	9	10^{-4}	24	7
`Cancer`	569	30	3	10^{-4}	363	2
`Credit`	690	14	2	10^{-4}	456	2

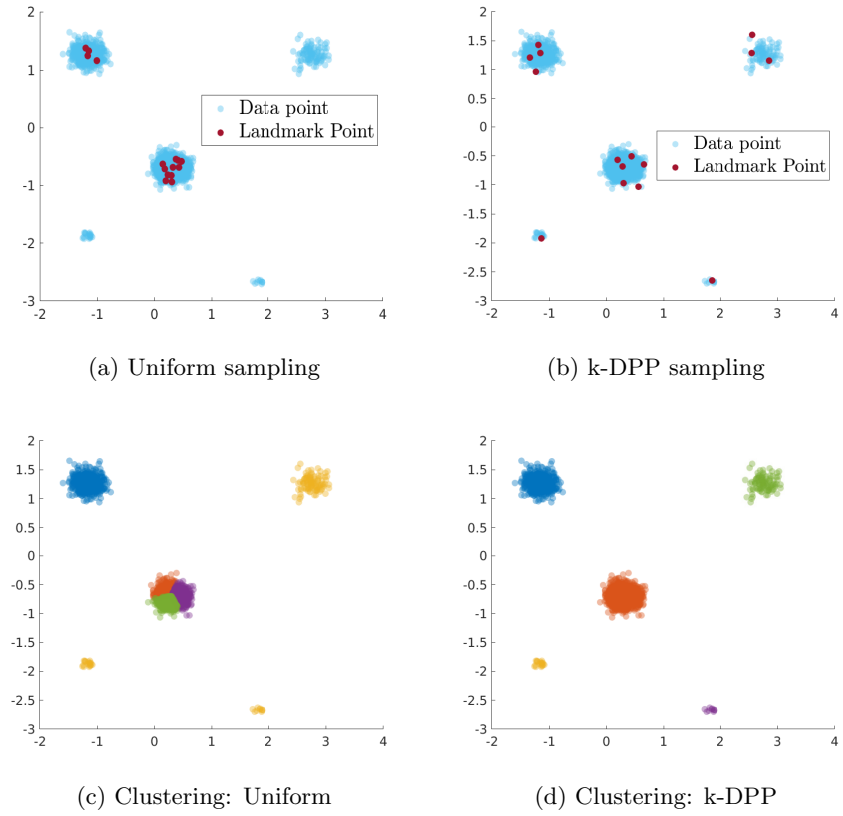
dataset consisting of 5 Gaussian bumps with a different number of points and different variances. Uniform sampling often only selects landmarks from 3 out of the 5 clusters, whereas a k-DPP samples landmarks out of every cluster. Consequently, kernel k-means algorithm with Nyström approximation can be improved by using diverse sampling. It is important to note that the superior performance is due to the

* **Funding:** * EU: The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.Research Council KUL: Optimization frameworks for deep kernel machines C14/18/068. Flemish Government:FWO: projects: GOA4917N (Deep Restricted Kernel Machines: Methods and Foundations), PhD/Postdoc grant.Impulsfonds AI: VR 2019 2203 DOC.0318/1QUATER Kenniscentrum Data en Maatschappij. Ford KU Leuven Research Alliance Project KUL0076 (Stability analysis and performance improvement of deep reinforcement learning algorithms). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders(FWO) and the Flemish Government – department EWI.

†KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. email: michael.fanuel@kuleuven.be

histogram of ridge-leverage scores with $\gamma = 10^{-4}$ (measure of outlyingness) having a long tail together with the different clusters being heavily imbalanced.

Next, we demonstrate the effect of diverse sampling on the `Glass`, `Breast Cancer` and `Australian Credit` datasets[1] of Table SM1. The condition number, smallest/largest eigenvalues of $K_{\mathcal{C}\mathcal{C}}$ and NMI is plotted as a function of the determinant. The averaged results are visualized in Figure SM2. Information on the datasets and hyperparameters used for the experiments is given in Table SM1. Sampling a diverse subset gives a more accurate clustering. Similar as for the kernel approximation experiments, in the presence of outliers, taking samples with an extremely large $\det(K_{\mathcal{C}\mathcal{C}})$ thanks to the Greedy Swapping Algorithm might decrease the accuracy as for the `Glass` dataset.



(a) Uniform sampling

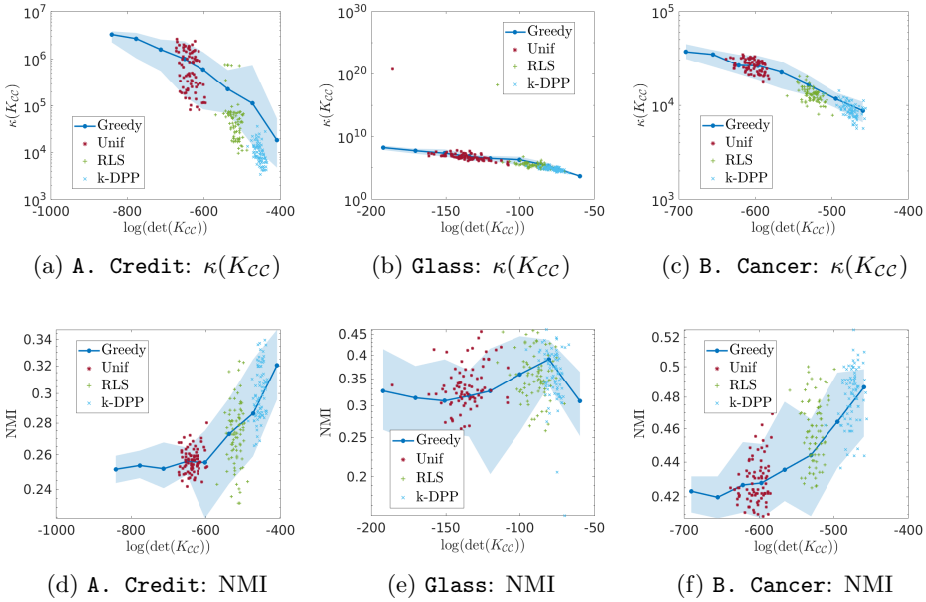
(b) k-DPP sampling

(c) Clustering: Uniform

(d) Clustering: k-DPP

Fig. SM1: Illustration of sampling methods on an artificial clustering problem. Uniform sampling oversamples dense parts, and does not select landmark points in every cluster. k-DPP sampling overcomes this limitation, resulting in better clustering performance.

SM0.2. Supplementary numerical experiments. Several additional illustrations, obtained with the main methodology as in the manuscript, on the datasets given in Table SM2 are given in the sequel. Nyström approximation error is given in Figures SM3 and SM8, Kernel PCA in Figure SM5, and Kernel Ridge Regression



(a) `A. Credit`: $\kappa(K_{CC})$ (b) `Glass`: $\kappa(K_{CC})$ (c) `B. Cancer`: $\kappa(K_{CC})$

(d) `A. Credit`: NMI (e) `Glass`: NMI (f) `B. Cancer`: NMI

Fig. SM2: Clustering results. The condition number and NMI are plotted as a function of $\log(\det(K_{CC}))$. A large NMI corresponds to a good accuracy.

in Figures SM6 and SM7. In most of the tasks illustrated in those figures, a larger diversity yields an improved performance. Let us discuss some particular cases.

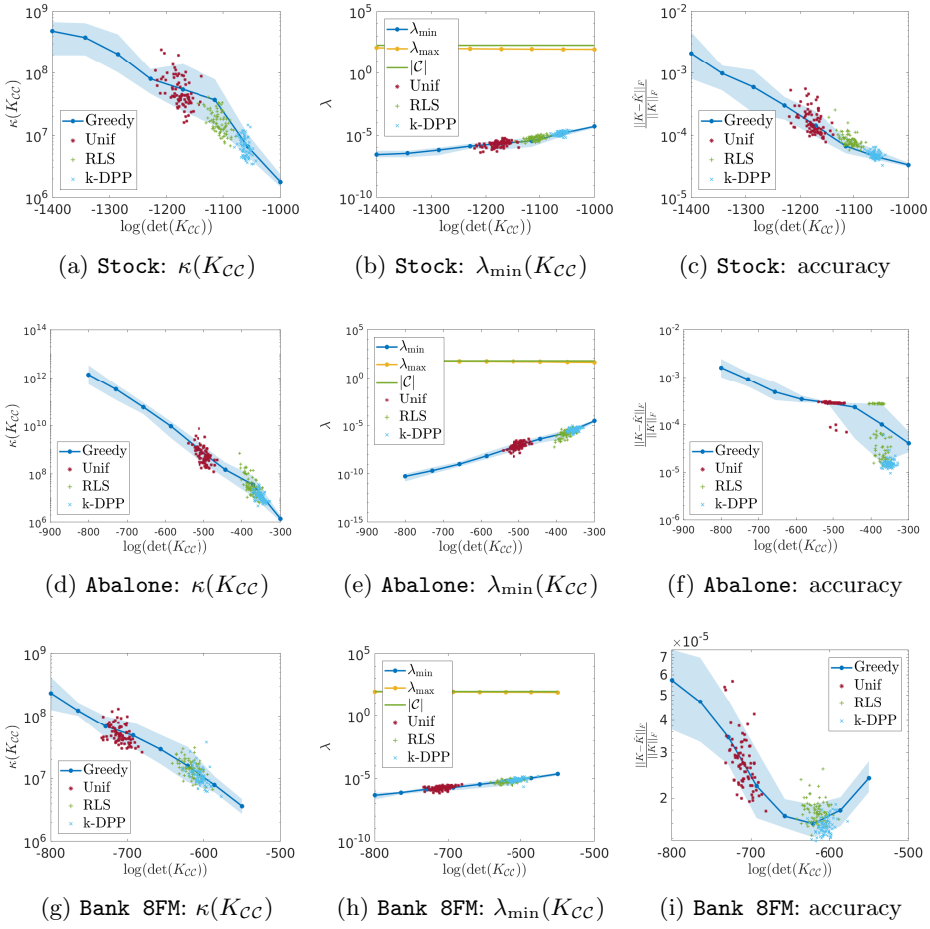
Kernel approximation. As it was mentioned already hereabove, in the presence of outliers, a very diverse subsample can produced a poor kernel approximation. This can be viewed in Figure SM3i, where the greedy algorithm is able to select subsets with a very large diversity. The randomized sampling methods that we studied empirically here did not suffer from this issue.

Table SM2: Datasets and parameters used for the experiments.

Dataset	Task	n	d	σ	λ	$d_{\text{eff}}(n\lambda)$
`Stock`	Kernel approx.	950	10	5	10^{-6}	119
`Abalone`	Kernel approx.	4177	8	10	10^{-6}	37
`Bank 8FM`	Kernel approx.	8192	8	10	10^{-6}	95
`Parkinson`	KPCA	5875	20	10	10^{-6}	219
`Wine Quality`	KPCA	6497	11	10	10^{-6}	133
`Housing`	KRR	506	13	3	10^{-4}	110
`Parkinson`	KRR	5875	20	3	10^{-4}	343

KRR. By using the same methodology as in the manuscript, the Mean Absolute Percentage Error is calculated both in the bulk and in the tail of the leverage score distribution of the test set. Notice that uniform sampling can often reduce the MAPE

in the bulk of the data, while diverse sampling yields a larger improvement in the tail of the distribution (cfr. Figure SM6e).

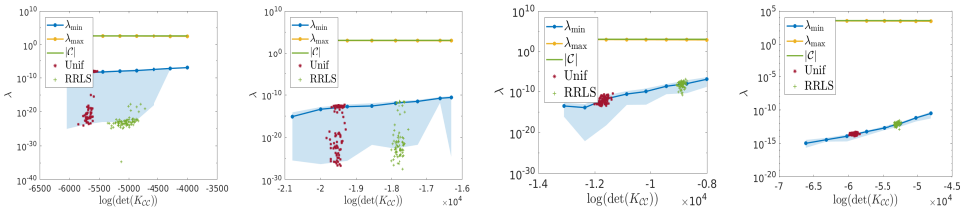


(a) `Stock`: $\kappa(K_{\mathcal{CC}})$ (b) `Stock`: $\lambda_{\min}(K_{\mathcal{CC}})$ (c) `Stock`: accuracy

(d) `Abalone`: $\kappa(K_{\mathcal{CC}})$ (e) `Abalone`: $\lambda_{\min}(K_{\mathcal{CC}})$ (f) `Abalone`: accuracy

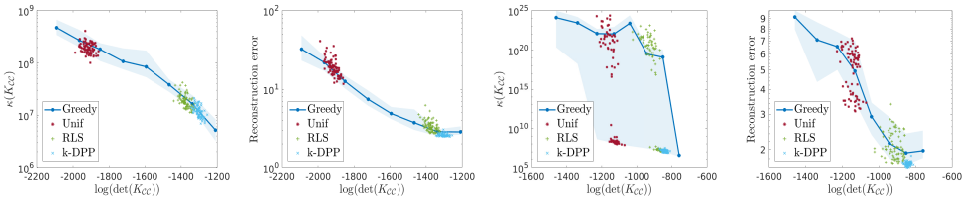
(g) `Bank 8FM`: $\kappa(K_{\mathcal{CC}})$ (h) `Bank 8FM`: $\lambda_{\min}(K_{\mathcal{CC}})$ (i) `Bank 8FM`: accuracy

Fig. SM3: Kernel approximation results. The condition number, smallest and largest eigenvalues of $K_{\mathcal{CC}}$, relative Frobenius norm of the Nyström approximation error versus $\log(\det(K_{\mathcal{CC}}))$.



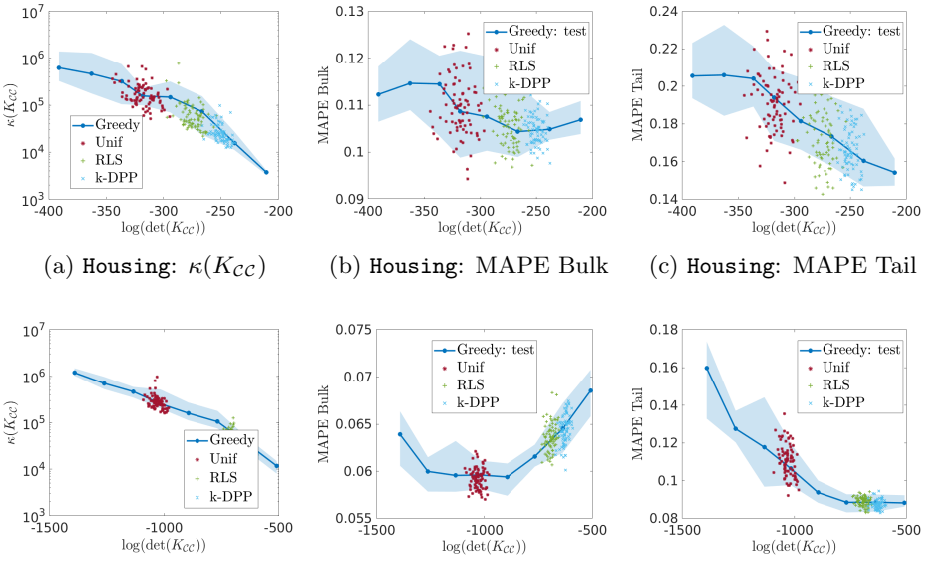
(a) `MiniBooNE` (b) `codRNA` (c) `Adult` (d) `Covertype`

Fig. SM4: Results accompanying the large-scale Nyström and KPCA experiments. The smallest and largest eigenvalues of $K_{\mathcal{CC}}$ are plotted as a function of $\log(\det(K_{\mathcal{CC}}))$.



(a) `Park.`: $\kappa(K_{\mathcal{CC}})$ (b) `Park.`: error (c) `Wine Q.`: $\kappa(K_{\mathcal{CC}})$ (d) `Wine Q.`: error

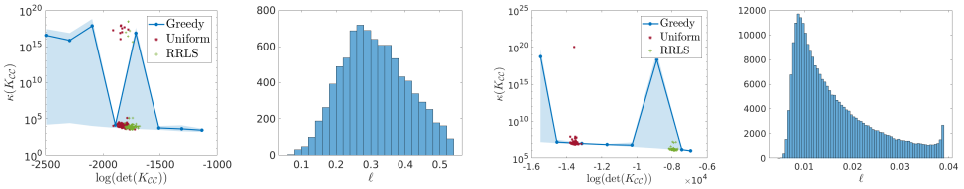
Fig. SM5: KPCA results. The condition number and reconstruction error using half of the components are plotted as a function of $\log(\det(K_{\mathcal{CC}}))$.



(a) `Housing`: $\kappa(K_{\mathcal{CC}})$ (b) `Housing`: MAPE Bulk (c) `Housing`: MAPE Tail

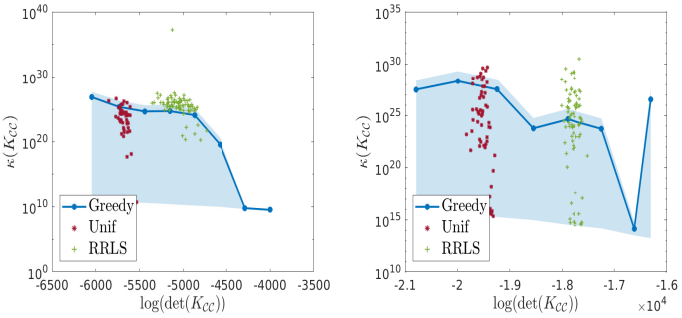
(d) `Park.`: $\kappa(K_{\mathcal{CC}})$ (e) `Park.`: MAPE Bulk (f) `Park.`: MAPE Tail

Fig. SM6: Regression results. The condition number and MAPE on the test set are plotted as a function of $\log(\det(K_{\mathcal{CC}}))$.



(a) `Bike S.`: $\kappa(K_{\mathcal{CC}})$ (b) `Bike S.`: RLS (c) `Year P.`: $\kappa(K_{\mathcal{CC}})$ (d) `Year P.`: RLS

Fig. SM7: Results accompanying the large-scale regression experiments. The condition number as a function of $\log(\det(K_{\mathcal{CC}}))$ and the approximate RLS distribution are visualized.



(a) `MiniBooNE`:$\kappa(K_{\mathcal{CC}})$ (b) `codRNA`: $\kappa(K_{\mathcal{CC}})$

Fig. SM8: Results accompanying the large-scale Nyström experiments. The condition number of $K_{\mathcal{CC}}$ versus the sample diversity.

REFERENCES

[SM1] A. Strehl and J. Ghosh, *Cluster ensembles—a knowledge reuse framework for combining multiple partitions*, Journal of machine learning research, 3 (2002), pp. 583–617.

[SM2] S. Wang, A. Gittens, and M. W. Mahoney, *Scalable kernel K-means clustering with Nyström approximation: relative-error bounds*, The Journal of Machine Learning Research, 20 (2019), pp. 431–479.