Towards Deterministic Diverse Subset Sampling

J. Schreurs[1], M. Fanuel[1], and J.A.K. Suykens[1]

KU Leuven, Department of Electrical Engineering (ESAT),
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics,
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{joachim.schreurs,michael.fanuel,johan.suykens}@kuleuven.be

Abstract. Determinantal point processes (DPPs) are well known models for diverse subset selection problems, including recommendation tasks, document summarization and image search. In this paper, we discuss a greedy deterministic adaptation of k-DPP. Deterministic algorithms are interesting for many applications, as they provide interpretability to the user by having no failure probability and always returning the same results. First, the ability of the method to yield low-rank approximations of kernel matrices is evaluated by comparing the accuracy of the Nyström approximation on multiple datasets. Afterwards, we demonstrate the usefulness of the model on an image search task.

Keywords: Determinantal point processes · Landmark sampling · Diversity.

1 Introduction

Selecting a diverse subset is an interesting problem for many applications. Examples are document or video summarization [3,10,13,14], image search tasks [11], pose estimation [13] and many others. Diverse sampling algorithms have also shown their benefits to calculate a low-rank matrix approximations using the Nyström method [21]. This method is a popular tool for scaling up kernel methods, where the quality of the approximation relies on selecting a representative subset of landmark points or Nyström centers.

Notations In this work, we will use uppercase letters for matrices and calligraphic letters for sets, while bold letter denote random variables. The notation $(\cdot)^\dagger$ denotes Moore-Penrose pseudo inverse of a matrix. We also define the partial order of positive definite (resp. semidefinite) matrices by $A \succ B$ (resp. $A \succeq B$) if and only if $A-B$ is positive definite (resp. semidefinite). Furthermore, we denote by K, the Gram matrix $[k(x_i, x_j)]_{i,j=1}^n$ obtained from a positive semidefinite kernel such as the Gaussian kernel $k(x, y) = \exp(-\|x - y\|_2^2/(2\sigma^2))$.

Nyström approximation The Nyström method takes a positive semidefinite matrix $K \in \mathbb{R}^{n\times n}$ as input, selects from it a small subset $\mathcal{C}$ of columns, and constructs the approximation $\hat{K} = K_{\mathcal{C}} K_{\mathcal{C}\mathcal{C}}^\dagger K_{\mathcal{C}}^\top$, where $K_{\mathcal{C}} = KC$ and $K_{\mathcal{C}\mathcal{C}} = C^\top KC$

are submatrices of the kernel matrix and $C \in \mathbb{R}^{n \times |\mathcal{C}|}$ is a sampling matrix obtained by selecting the columns of the identity matrix indexed by $\mathcal{C}$. The matrix $\hat{K}$ is used in the place of K, so to decrease the training runtime and memory requirements. Using a dependent or *diverse* sampling algorithm for the Nyström approximation has shown to give better performance than independent sampling methods in [8,15].

Determinantal Point Processes and kernel methods Determinantal point processes (DPPs) [12] are well known models for diverse subset selection problems. A point process on a ground set $[n] = \{1, 2, ..., n\}$ is a probability measure over point patterns, which are finite subsets of $[n]$. It is common to define a DPP thanks to its marginal kernel, that is a positive symmetric semidefinite matrix satisfying $P \preceq \mathbb{I}$. Let $\boldsymbol{\mathcal{Y}}$ denote a random subset, drawn according to the DPP with marginal kernel P. Then, the probability that $\mathcal{C}$ is a subset of the random $\boldsymbol{\mathcal{Y}}$ is defined by

$$\Pr(\mathcal{C} \subseteq \boldsymbol{\mathcal{Y}}) = \det(P_{\mathcal{C}\mathcal{C}}). \tag{1}$$

Notice that all principal submatrices of a positive semidefinite matrix are positive semidefinite. From (1), it follows that:

$$\begin{aligned} \Pr(i \in \boldsymbol{\mathcal{Y}}) &= P_{ii} \\ \Pr(i, j \in \boldsymbol{\mathcal{Y}}) &= P_{ii}P_{jj} - P_{ij}P_{ji} \\ &= \Pr(i \in \boldsymbol{\mathcal{Y}})\Pr(j \in \boldsymbol{\mathcal{Y}}) - P_{ij}^2. \end{aligned}$$

The diagonal elements of the kernel matrix give the marginal probability of inclusion for individual elements, whereas the off-diagonal elements determine the "repulsion" between pairs of elements. Thus, for large values of P_{ij}, or a high similarity, points are unlikely to appear together. In some applications, it can be more convenient to define DPPs thanks to L-ensembles, which can be related to marginal kernels by the formula $L = P(\mathbb{I} - P)^{-1}$ when $P \prec \mathbb{I}$ (more details in [2]). They allow to define the probability of sampling a random subset $\boldsymbol{\mathcal{Y}}$ that is equal to $\mathcal{C}$:

$$\Pr(\boldsymbol{\mathcal{Y}} = \mathcal{C}) = \frac{\det(L_{\mathcal{C}\mathcal{C}})}{\det(\mathbb{I} + L)}. \tag{2}$$

In contrast to (1), the only requirement on L is that it has to be positive semidefinite. Notice that the normalization in (2) can be derived classically by considering the property relating the coefficients of the characteristic polynomial of a matrix to the sum of the determinant of its principal submatrices of the same size. In this paper, the L-ensemble is chosen to be a kernel matrix K.

Exact sampling of a DPP is done in two phases [12]. Let V be the matrix whose columns are the eigenvectors of K. In the first phase, a subset of eigenvectors of the kernel matrix K is selected at random, where the probability of selecting each eigenvector depends on its associated eigenvalue in a specific way given in Algorithm 1. In the second phase, a sample $\mathcal{Y}$ is produced based on the selected vectors. At each iteration of the second loop, the cardinality of $\mathcal{Y}$ increases by one and the number of columns of V is reduced by one. A k-DPP [11]

is a DPP conditioned on a fixed cardinality $|\mathcal{Y}| = k$. Note that $e_i \in \mathbb{R}^n$ is the i-th vector of the canonical basis.

input: L-ensemble $L \succeq 0$
initialization: $\mathcal{J} = \emptyset$ and $\mathcal{Y} = \emptyset$
Calculate the eigenvector/value pairs $\{(v_i, \lambda_i)\}_{i=1}^n$ of L.
for: $i = 1, \ldots, n$ **do**
  $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$ with prob. $\frac{\lambda_i}{\lambda_i + 1}$
end for
$V \leftarrow \{v_i\}_{i \in \mathcal{J}}$ a set of columns
while: $|V| > 0$ **do**
  Draw an index i according to the distribution $p_i = \frac{1}{|V|} \sum_{v \in V} (v^{\mathrm{T}} e_i)^2$.
  $\mathcal{Y} \leftarrow \mathcal{Y} \cup i$
  $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to e_i.
end while
return $\mathcal{Y}$.

Algorithm 1: Exact DPP sampling algorithm associated to the L-ensemble L [12]. Notice that $P = L(L + \mathbb{I})^{-1}$, so that the eigenvector/value pairs of P are exactly $\{(v_i, \frac{\lambda_i}{\lambda_i + 1})\}_{i=1}^n$.

Deterministic algorithms are interesting for many applications, as they provide interpretability to the user by having no chance of failure and always returning the same results. The usefulness of deterministic algorithms has already been recognized by Papailiopoulos et al. [18] and McCurdy [17], who provide deterministic algorithms based on the (ridge) leverage scores. These statistical leverage scores correspond to correlations between the singular vectors of a matrix and the canonical basis [1,7]. The recently introduced Deterministic Adaptive Sampling (DAS) algorithm [8] provides a deterministically obtained *diverse* subset. The method shows superior performance compared to randomized counterparts in terms of approximation error for the Nyström approximation when the eigenvalues of the kernel matrix have a fast decay. A similar observation was made for the deterministic algorithms of Papailiopoulos et al. [18] and McCurdy [17].

This paper discusses a deterministic adaption of k-DPP, where we have the following empirical observations:

1. The method is deterministic, hence there is no failure probability and the method always produces the same output.
2. Only the k eigenvectors with the largest eigenvalues are needed, which results in a speedup when $k \ll n$.
3. We observed that the method samples a more diverse subset than the original k-DPP on multiple datasets.
4. There is no need to tune a regularization parameter, which is the case for the DAS algorithm.

5. The method shows superior accuracy in terms of the max norm of the Nyström approximation on multiple datasets compared to the standard k-DPP, along with better accuracy of the kernel approximation for the operator norm when there is fast decay of the eigenvalues.

In Section 2, we introduce the method. Secondly, we make a connection with the DAS algorithm, namely the deterministic k-DPP corresponds to the DAS algorithm with an adapted projector kernel matrix. In Section 3, we evaluate the method on different datasets. Finally, a small real-life illustration is shown in Section 4.

2 Deterministic adaptation of k-DPP

We discuss a deterministic adaptation of k-DPP, by selecting iteratively landmarks with the highest probability. As it is described in Algorithm 2, we can successively maximize the probability over a nested sequence of sets $\mathcal{C}_0 \subseteq \mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_k$ starting with $\mathcal{C}_0 = \emptyset$ by adding one landmark at each iteration. The proposed method is an adaptation of the improved k-DPP sampling algorithm given by Tremblay et al. [20]. Our proposed method start from a projective marginal kernel $P = VV^\top$, with $V = [v_1, ..., v_k] \in \mathbb{R}^{n\times k}$ the sampled eigenvectors of the kernel matrix. Instead of sampling the eigenvectors [12], the k eigenvectors with the largest eigenvalue are chosen. Secondly, at each iteration the point with the highest probability $p(i) = P_{ii} - P_{\mathcal{C}i}^\top P_{\mathcal{C}\mathcal{C}}^\dagger P_{\mathcal{C}i}$ is chosen, where $\mathcal{C}$ corresponds to the selected subset so far. Besides the interpretation of DPPs in relation to diversity, the aforementioned probability gives a second insight in diversity. Namely we have $p(i) = \|v_i - \pi_{V_\mathcal{C}} v_i\|_2^2$, where $v_i \in \mathbb{R}^{n\times 1}$ is the i-th column of V and $\pi_{V_\mathcal{C}}$ is the projector on $V_\mathcal{C} = \operatorname{span}\{v_s | s \in \mathcal{C}\}$. The chosen landmark corresponds to the point that is the most distant to the space of the previously sampled points.

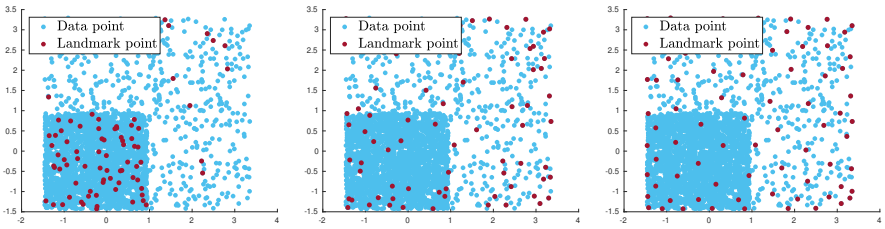
input: Kernel matrix K, sample size k.
initialization: $\mathcal{C} \leftarrow \emptyset$
Calculate the first k eigenvectors $\mathrm{V} \in \mathbb{R}^{n\times k}$ from K.
$P = VV^{\mathrm{T}}$
Define $p_0 \in \mathbb{R}^N : \forall i, \quad p_0(i) = \left\|\mathrm{V}^\top e_i\right\|^2$
$p \leftarrow p_0$
for: $i = 1, \ldots, k$ **do**
  Select c_i with highest probability $p(i)$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$
  Update $\boldsymbol{p} : \forall j \quad p(j) = p_0(j) - P_{\mathcal{C}j}^\top P_{\mathcal{C}\mathcal{C}}^\dagger P_{\mathcal{C}j}$
end for
return $\mathcal{C}$.

Algorithm 2: Deterministic adaptation of the k-DPP sampling algorithm.

2.1 Connections with DAS

Algorithmically, the proposed method corresponds to the DAS algorithm [8] (see, Algorithm 3 in appendix) with a different projector kernel matrix. More precisely, DAS uses a smoothed projector kernel matrix $P_{n\gamma}(K) = K(K+n\gamma I)^{-1}$ with $K \succ 0$. The ridge leverage scores $l_i(\gamma) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j+n\gamma} V_{ij}^2$ can be found on the diagonal $P_{n\gamma}(K)$. Let $V \in \mathbb{R}^{n\times n}$ be the matrix of eigenvectors of the kernel matrix K. On the contrary, in this paper, the proposed method has a *sharp projector kernel matrix* with the rank-k leverage scores $l_i = \sum_{j=1}^{k} V_{ij}^2$ on the diagonal. This has the added benefit that there is no regularization parameter to tune. The DAS algorithm is given in the Appendix.

Algorithm 2 is a greedy reduced basis method as defined in [6]. Other greedy maximum volume approaches are found in [4,5,9]. These greedy methods are used for finding an estimate of the most likely configuration (MAP), which is known to be NP-hard [9]. In practice, we see that the method performs quite well and gives a consistently larger $\det(K_{\mathcal{CC}})$, which is considered a measure for diversity, compared to the randomized counterpart (see Section 3). This number is calculated as follows $\log(\det(K_{\mathcal{CC}})) = \sum_{i=1}^{k} \log(\lambda_i)$, with $\{\lambda_i\}_{i=1}^{k}$ the singular values of $K_{\mathcal{CC}}$. A small illustration is given in Figure 1, where the deterministic algorithm gives a more diverse subset.



(a) Uniform (b) k-DPP (c) Deterministic k-DPP

Fig. 1: Illustration of sampling methods on an artificial dataset. Uniform sampling does not promote diversity and selects almost all points in the bulk of the data. Sampling a k-DPP overcomes this limitation, however landmarks can be close to each other. The latter is solved by using the deterministic adaptation of k-DPP, which gives a more diverse subset.

3 Numerical results

We evaluate the performance of the deterministic variant of the k-DPP with a Gaussian kernel on the `Boston housing`, `Stock`, `Abalone` and `Bank 8FM` datasets,

which have 506, 950, 4177 and 8192 datapoints respectively. Those public datasets[1] have been used for benchmarking k-DPPs in [15]. The implementation of the algorithms is done with MatlabR2018b.

Throughout the experiments, we use a fixed bandwidth $\sigma = 2$ for `Boston housing`, `Stock` and `Abalone` dataset and $\sigma = 5$ for the `Bank 8FM` dataset after standardizing the data. The following algorithms are used to sample k landmarks: Uniform sampling, k-DPP[2] [11], DAS [8] and the proposed method. DAS is executed for multiple regularization parameters $\gamma \in \{10^0, 10^{-1}, \ldots, 10^{-6}\}$ where the sample with the best performing γ is selected to approximate the kernel matrix. The total experiment is repeated 10 times. The quality of the landmarks $\mathcal{C}$ is evaluated by the relative operator or max norm $\|K - \hat{K}\|_{\{\infty,2\}} / \|K\|_{\{\infty,2\}}$ with $\hat{K} = K_{\mathcal{C}}(K_{\mathcal{C}\mathcal{C}} + \epsilon \mathbb{I})^{-1} K_{\mathcal{C}}^{\top}$ with $\varepsilon = 10^{-12}$ for numerical stability. 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 diversity is measured by $\log(\det(K_{\mathcal{C}\mathcal{C}}))$, where a larger log determinant means more diversity. The results for the `Stock` dataset are visible in Figure 2. The results for the rest of the datasets or shown in Figures 8, 9, 10 and 11 in the Appendix. The computer used for these simulations has 8 processors $3.40GHz$ and 15.5 GB of RAM.

As previously mentioned, the greedy method returns a more diverse subset. Figure 8 shows the $\log(\det(K_{\mathcal{C}\mathcal{C}}))$ where the proposed method shows similar performance as DAS, while improving on both the randomized methods. The same is visible for the relative max norm of the Nyström approximation error. DAS and the deterministic variant of the k-DPP perform well on the `Boston housing` and `Stock` dataset, which show a fast decay in the spectrum of K (see Figure 7). If the decay of the eigenvalues is not fast enough, the randomized k-DPP, shows better performance. The same observation was made for the deterministic (ridge) leverage score sampling algorithms [17,18] as well as DAS [8].

4 Illustration

We demonstrate the use of the proposed method on a image summarization task. The first experiment is done on the `Stanford Dogs` dataset[3], which contains images of 120 breeds of dogs from around the world. This dataset has been built using images and annotation from ImageNet for the task of fine-grained image categorization. The training features are SIFT descriptors [16] (given by the dataset), which are used to make a histogram intersection kernel. We take a subset of 50 images each of the classes *border collie*, *chihuahua* and *golden retriever*. The total training set is visualized in Figure 5 in the Appendix. Figure 3 displays the results of the method for $k = 4$. One can observe that the images are very dissimilar and dogs out of each breed are represented. This is confirmed by the projection of the landmarks on the 2 first principal components of the

[1] `https://www.cs.toronto.edu/~delve/data/datasets.html`, `https://www.openml.org/d/223`

[2] We used the Matlab code available at `https://www.alexkulesza.com/`.

[3] `http://vision.stanford.edu/aditya86/ImageNetDogs/main.html`

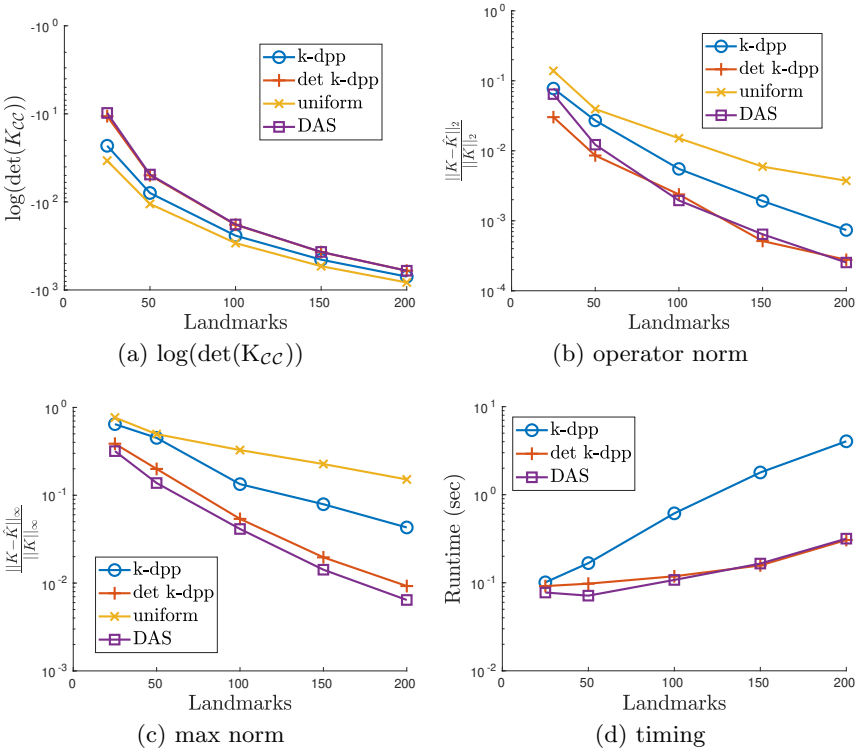


Fig. 2: The $\log(\det(K_{CC}))$, relative operator norm and relative max norm of the Nyström approximation error and timings as a function of the number of landmarks on the Stock dataset. The results are plotted on a logarithmic scale, averaged over 10 trials. The larger $\log(\det(K_{CC}))$, the more diverse the subset.

kernel principal component analysis (KPCA) [19], where the landmark points lie in the outer regions of the space.

We repeat the above procedure on the `Kimia99` dataset[4]. The dataset has 9 classes consisting of 11 images each. It contains shapes silhouettes for the classes: rabbits, quadrupeds, men, airplanes, fish, hands, rays, tools, and a miscellaneous class. The total training set is visible in Figure 6 in the Appendix. First, we resize the images to size 100×100. Afterwards, we apply a Gaussian kernel with bandwidth $\sigma = 100$ after standardizing the data. The results of the proposed method with $k = 9$ are visible in Figure 4. The method samples landmarks out every class, making it a desirable image summarization. This is supported by the projection of the landmarks on the 2 first principal components of the KPCA, where a landmark points is chosen out of every small cluster.

[4] `https://vision.lems.brown.edu/content/available-software-and-databases#Datasets-Shape`

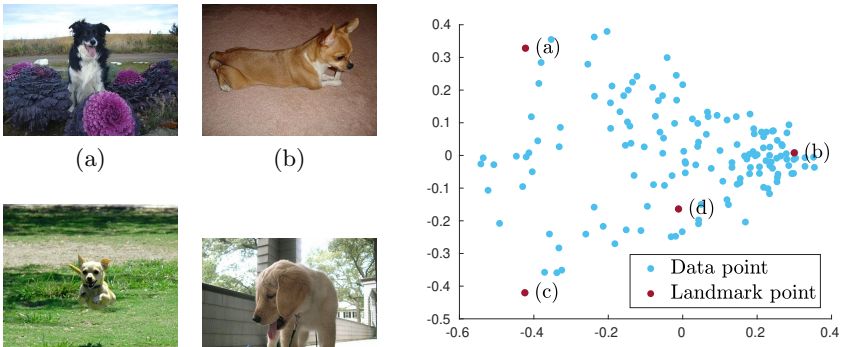


Fig. 3: Illustration of the proposed method with $k = 4$ on the `Stanford Dogs` dataset. The selected landmark points are visualized on the left, the projection on the 2 first principal components of the KPCA on the right.

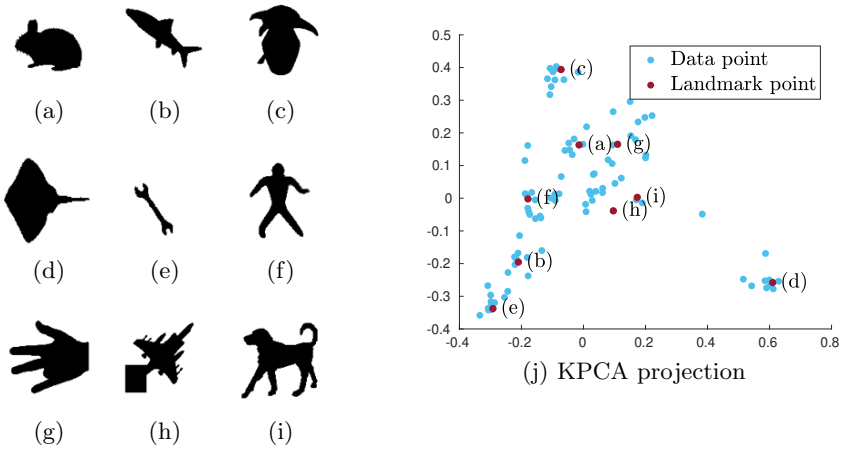


Fig. 4: Illustration of the proposed method with $k = 9$ on the `Kimia99` dataset. The selected landmark points are visualized on the left, the projection on the 2 first principal components of the KPCA on the right.

5 Conclusion

We discussed a greedy deterministic adaptation of k-DPP. Algorithmically, the method corresponds to the DAS algorithm with a different projector kernel matrix. The proposed method is evaluated by comparing the accuracy of the

Nyström approximation on multiple datasets. Experiments show the proposed method is able to give a more diverse subset, along with better performance for the relative max norm. When there is a fast decay of the eigenvalues, the deterministic method is more accurate than randomized counterparts. To conclude, we demonstrate the usefulness of the model on an image search task.

Acknowledgements

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).

References

1. Alaoui, A., Mahoney, M.W.: Fast randomized kernel ridge regression with statistical guarantees. In: Advances in Neural Information Processing Systems. pp. 775–783 (2015)
2. Borodin, A.: Determinantal point processes. arXiv preprint arXiv:0911.1153 (2009)
3. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. vol. 98, pp. 335–336 (1998)
4. Chen, L., Zhang, G., Zhou, E.: Fast greedy map inference for determinantal point process to improve recommendation diversity. In: Advances in Neural Information Processing Systems. pp. 5622–5633 (2018)
5. Çivril, A., Magdon-Ismail, M.: On selecting a maximum volume sub-matrix of a matrix and related problems. Theoretical Computer Science **410**(47-49), 4801–4811 (2009)
6. DeVore, R., Petrova, G., Wojtaszczyk, P.: Greedy algorithms for reduced bases in banach spaces. Constructive Approximation **37**(3), 455–466 (2013)
7. Drineas, P., Magdon-Ismail, M., Mahoney, M.W., Woodruff, D.P.: Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research **13**(Dec), 3475–3506 (2012)
8. Fanuel, M., Schreurs, J., Suykens, J.A.K.: Nyström landmark sampling and regularized Christoffel functions. arXiv preprint arXiv:1905.12346 (2019)
9. Gillenwater, J., Kulesza, A., Taskar, B.: Near-optimal map inference for determinantal point processes. In: Advances in Neural Information Processing Systems. pp. 2735–2743 (2012)
10. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Advances in Neural Information Processing Systems. pp. 2069–2077 (2014)
11. Kulesza, A., Taskar, B.: k-dpps: Fixed-size determinantal point processes. In: Proceedings of the 28th International Conference on Machine Learning. pp. 1193–1200 (2011)
12. Kulesza, A., Taskar, B.: Determinantal point processes for machine learning. Foundations and Trends in Machine Learning **5**(2-3), 123–286 (2012)
13. Kulesza, A., Taskar, B.: Structured determinantal point processes. In: Advances in neural information processing systems. pp. 1171–1179 (2010)
14. Kulesza, A., Taskar, B.: Learning determinantal point processes. arXiv preprint arXiv:1202.3738 (2012)
15. Li, C., Jegelka, S., Sra, S.: Fast DPP sampling for Nyström with application to kernel methods. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning. pp. 2061–2070 (2016)
16. Lowe, D.G., et al.: Object recognition from local scale-invariant features. In: ICCV. vol. 99, pp. 1150–1157 (1999)
17. McCurdy, S.: Ridge regression and provable deterministic ridge leverage score sampling. In: Advances in Neural Information Processing Systems 31, pp. 2468–2477 (2018)
18. Papailiopoulos, D., Kyrillidis, A., Boutsidis, C.: Provable deterministic leverage score sampling. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 997–1006 (2014)
19. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: International conference on artificial neural networks. pp. 583–588. Springer (1997)

20. Tremblay, N., Barthelme, S., Amblard, P.O.: Optimized algorithms to sample determinantal point processes. arXiv preprint arXiv:1802.08471 (2018)
21. Williams, C.K., Seeger, M.: Using the Nyström method to speed up kernel machines. In: Advances in neural information processing systems. pp. 682–688 (2001)

A Additional algorithms

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

Algorithm 3: DAS algorithm [8].

B Additional figures

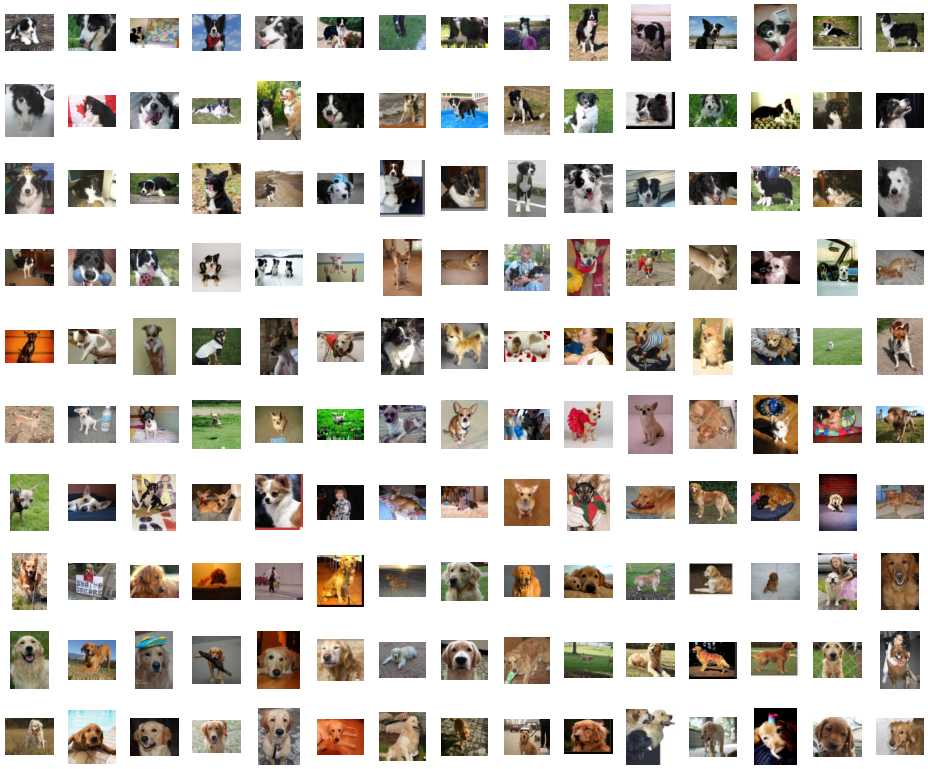


Fig. 5: The training data of the `Stanford Dogs` dataset.

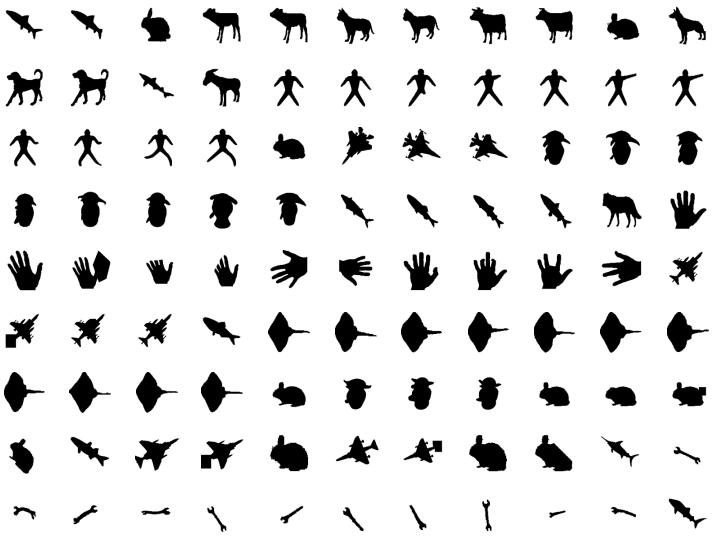
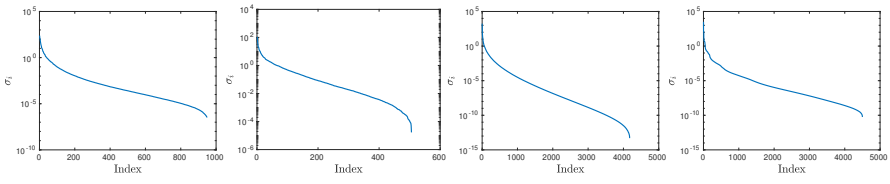
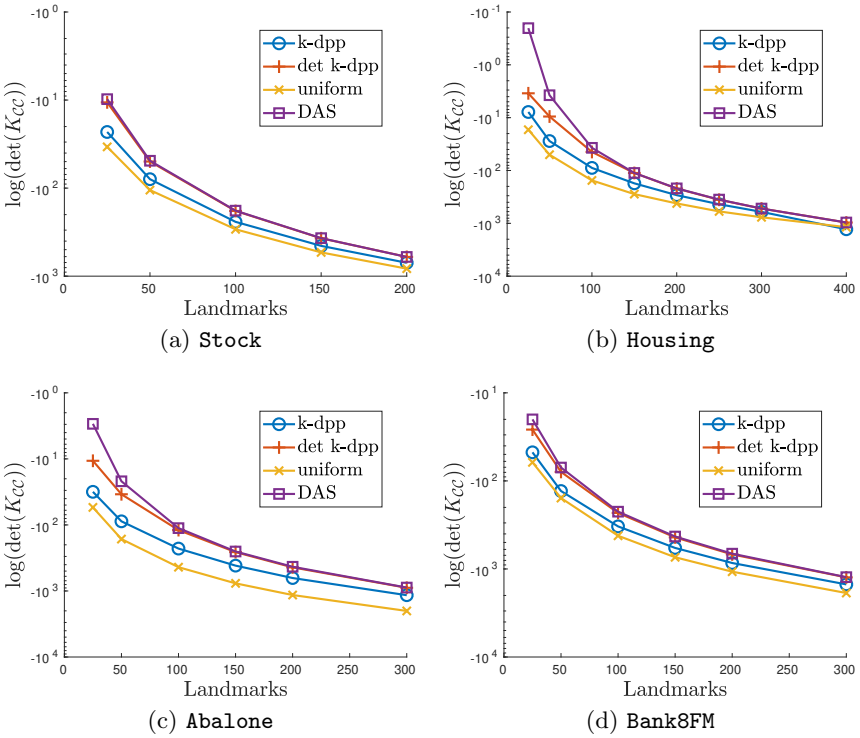


Fig. 6: The training data of the `Kimia99` dataset.



(a) `Stock` (b) `Housing` (c) `Abalone` (d) `Bank8FM`

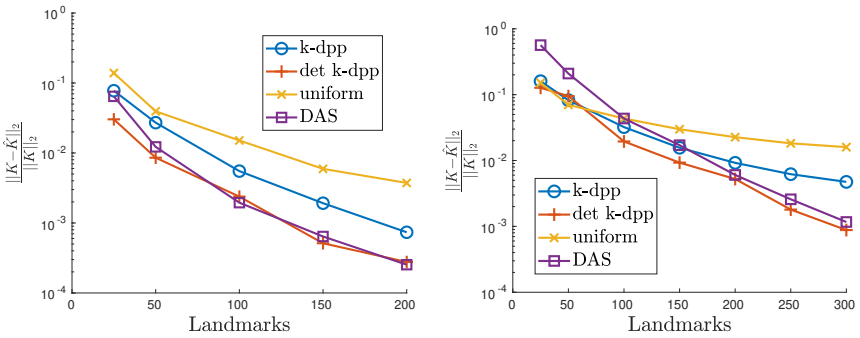
Fig. 7: Singular value spectrum of the datasets on a logarithmic scale. For a given index, the value of the eigenvalues for the `Stock` and `Housing` dataset are smaller than `Abalone` and `Bank8FM`.



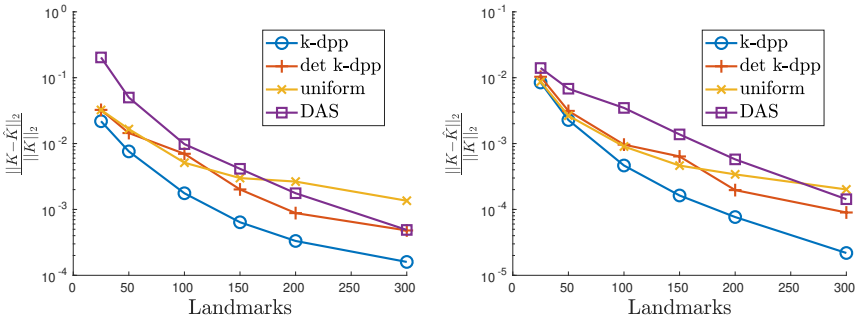
(a) `Stock` (b) `Housing`

(c) `Abalone` (d) `Bank8FM`

Fig. 8: $\log(\det(K_{\mathcal{CC}}))$ in function of the number of landmarks. The error is plotted on a logarithmic scale, averaged over 10 trials. The larger the $\log(\det(K_{\mathcal{CC}}))$, the more diverse the subset.

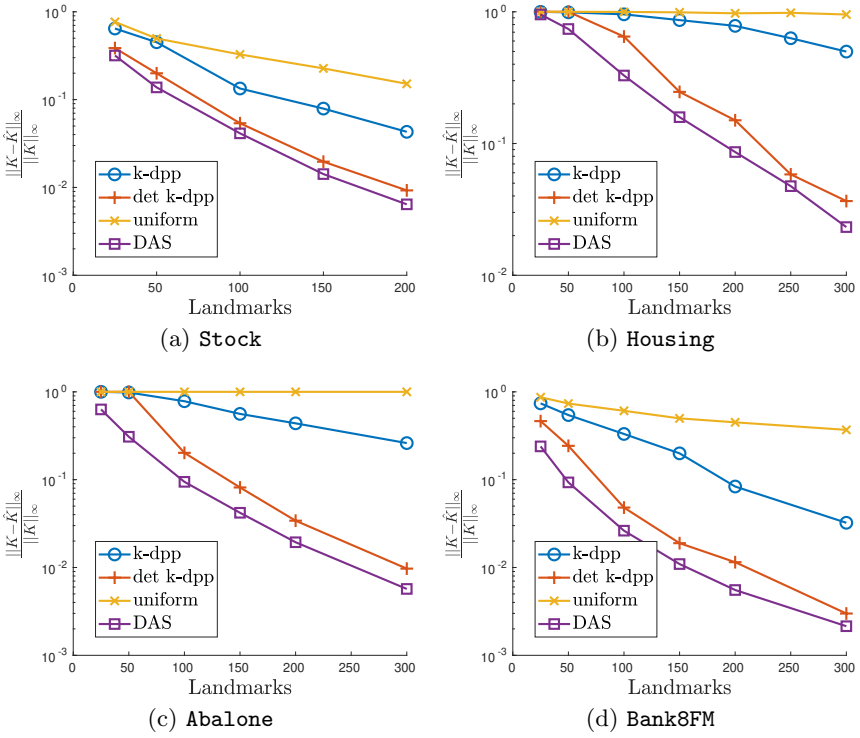


(a) `Stock` (b) `Housing`



(c) `Abalone` (d) `Bank8FM`

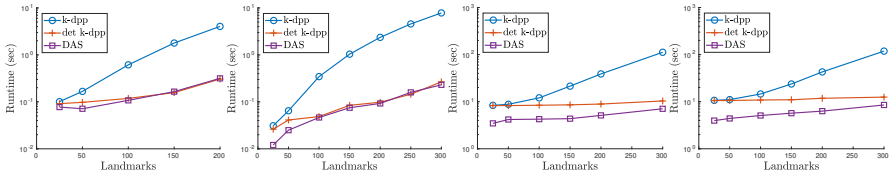
Fig. 9: Relative operator norm of the Nyström approximation error as a function of the number of landmarks. The error is plotted on a logarithmic scale, averaged over 10 trials.



(a) `Stock` (b) `Housing`

(c) `Abalone` (d) `Bank8FM`

Fig. 10: Relative max norm 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. 11: Timings for the computations of Figure 9 as a function of the number of landmarks. The timings are plotted on a logarithmic scale, averaged over 10 trials.