Ensemble Kernel Methods, Implicit Regularization and Determinantal Point Processes

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

Abstract

By using the framework of Determinantal Point Processes (DPPs), some theoretical results concerning the interplay between diversity and regularization can be obtained. In this paper we show that sampling subsets with kDPPs results in implicit regularization in the context of ridgeless Kernel Regression. Furthermore, we leverage the common setup of state-of-the-art DPP algorithms to sample multiple small subsets and use them in an ensemble of ridgeless regressions. Our first empirical results indicate that ensemble of ridgeless regressors can be interesting to use for datasets including redundant information.

1. Introduction

Recent work has shown numerous insightful connections between Determinantal Point Processes (DPPs) and implicit regularization which lead to new guarantees and improved algorithms. The so-called DPPs are probabilistic models of repulsion inspired from physics, which are capable of sampling diverse subsets. An extensive overview of the use of DPPs in randomized linear algebra can be found in (Dereziński & Mahoney, 2020). By utilizing DPPs, exact expressions for implicit regularization as well as connections to the double descent curve (Belkin et al., 2019) were derived in (Fanuel et al., 2020; Dereziński et al., 2019; 2020). The nice theoretical properties of DPPs sparked the search for efficient sampling algorithms. This resulted in many alternative algorithms for DPPs to reduce the computational cost of preprocessing and/or sampling, including many approximate and heuristic approaches. Some examples are the exact sampler without eigendecomposition of (Desolneux et al., 2018; Poulson, 2020), coreDPP of (Li et al., 2016) or the DPP-VFX algorithm of (Dereziński et al., 2019). The computational cost is often split in two parts: an initial preprocessing cost and subsequent sampling cost. The latter is typically smaller, which makes the previously mentioned algorithms especially useful for sampling multiple small subsets from a large dataset.

We extend the work of (Fanuel et al., 2020), where the role of *diversity* within kernel methods was investigated. Namely, a more diverse subset results in implicit regularization, which in turn improves the performance of different kernel applicationsMore specifically we generalize the implicit regularization of DPPs to kDPPs, which are DPPs conditioned on a fixed subset size k (Kulesza & Taskar, 2011). Furthermore, we leverage the common setup of state of the art DPP sampling algorithms, to sample multiple small subsets and use them in an ensemble approach. One can loosely characterize these ensemble approaches as methods wherein the data points are divided into smaller subsets, and estimators are trained on the divisions. Their use has shown to improve performance in Nyström approximation (Kumar et al., 2009) and kernel ridge regression (Zhang et al., 2013; Hsieh et al., 2014; Lin et al., 2017). Experiments show a reduction in error when combining multiple diverse subsets.

Nyström Approximation. Let $k(x,y) > 0$ be a continuous and *strictly* positive definite kernel. Examples are the Gaussian kernel $k(x,y) = \exp(-\|x-y\|_2^2/2\sigma^2)$ or Laplace Kernel $k(x,y) = \exp(-\|x-y\|_2/\sigma)$. Given data $\{x_i \in \mathbb{R}^d\}_{i\in[n]}$, kernel methods rely on the entries of the Gram matrix $K = [k(x_i,x_j)]_{i,j}$. By assumption, this Gram matrix is invertible. However, to avoid inverting the full Gram matrix, one often samples a subset of landmarks $\mathcal{C} \subseteq [n]$ with $n \times |\mathcal{C}|$ a sampling matrix C obtained by selecting the columns of the identity matrix indexed by $\mathcal{C}$. Next we define: $K_{\mathcal{C}} = KC$ and $K_{\mathcal{CC}} = C^\top KC$. Then, the $n\times n$ kernel matrix K is approximated by a low rank Nyström approximation $L(K,\mathcal{C}) = K_{\mathcal{C}} K_{\mathcal{CC}}^{-1} K_{\mathcal{C}}^\top$, which involves inverting the smaller $K_{\mathcal{CC}}$.

Ridgeless Kernel Regression. Given input-output pairs $\{(x_i,y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i\in[n]}$, we propose to solve

$$f_{\mathcal{C}}^\star = \arg\min_{f\in\mathcal{H}} \|f\|_{\mathcal{H}}^2, \text{ s.t. } y_i = f(x_i) \text{ for all } i \in \mathcal{C}, \quad (1)$$

[1]Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Leuven, Belgium. Correspondence to: Joachim Schreurs <joachim.schreurs@kuleuven.be>.

ICML 2020 workshop on Negative Dependence and Submodularity for ML, Vienna, Austria, PMLR 119, 2020.

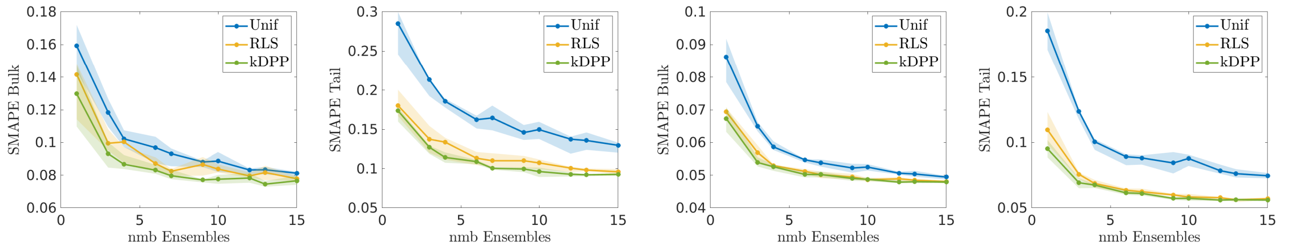

Figure 1. Ensemble KRR on the Abalone and Wine Quality dataset (from left to right).The SMAPE on the bulk and tail of the dataset is given in function of the number of ensembles.

where $\mathcal{C} \subseteq [n]$ is sampled by using a DPP. Here, $\mathcal{H}$ is the reproducing kernel Hibert space associated with k. The expression of the solution is $f^\star_{\mathcal{C}}(x) = \boldsymbol{k}_x^\top C K_{\mathcal{CC}}^{-1} C^\top \boldsymbol{y}$, where $\boldsymbol{k}_x = [k(x, x_1), \ldots, k(x, x_n)]^\top$. This approximation assumes that some data points can be omitted, contrary to Nyström approximation to Kernel Ridge Regression (KRR) which uses all data points. We show in this paper that averaging ridgeless regressors yield the solution of a regularized Kernel Ridge Regression calculated over the complete dataset. For $\mathcal{C} \sim DPP(K/\alpha)$, the expectation of the rigdeless predictors (cfr. Theorem 1) gives the function

$$\mathbb{E}_C[f^\star_{\mathcal{C}}(x)] = \boldsymbol{k}_x^\top (K + \alpha \mathbb{I})^{-1} \boldsymbol{y} =: f^\star(x) \tag{2}$$

which is the solution of Kernel Ridge Regression with a ridge parameter associated to α, namely

$$f^\star = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \|f\|^2_{\mathcal{H}}.$$

Typically, a large $\alpha > 0$ yields a small expected subset size for $DPP(K/\alpha)$. In light of the expectation result of (2), we propose to sample multiple subsets using a DPP and average the ridgeless predictors in an ensemble approach: $\bar{f} = \frac{1}{m} \sum_{i=1}^{m} f^\star_{\mathcal{C}_i}$ with m the number of ensembles.

Determinantal Point Processes A more extensive overview of DPPs is given in (Kulesza & Taskar, 2012). Let L be a $n \times n$ positive definite symmetric matrix, called L-ensemble. The probability of sampling a subset $\mathcal{C} \subseteq [n]$ is defined as follows $\Pr(Y = \mathcal{C}) = \det(L_{\mathcal{CC}}) / \det(\mathbb{I} + L)$. Where we define $L = K/\alpha$ with $\alpha > 0$ and denote the associated process $DPP_L(K/\alpha)$. The inclusion probabilities are given by $\Pr(\mathcal{C} \subseteq Y) = \det(P_{\mathcal{CC}})$, where $P = K(K + \alpha \mathbb{I})^{-1}$, is the marginal kernel associated to the L-ensemble $L = K/\alpha$. The diagonal of this soft projector matrix P gives the Ridge Leverage Scores (RLS) of the data points: $\boldsymbol{\ell}_i = P_{ii}$ for $i \in [n]$, which have been used to sample landmarks points in various works (Bach, 2013; El Alaoui & Mahoney, 2015; Musco & Musco, 2017) in the context of Nyström approximations. The RLS can be viewed as the importance or uniqueness of a data point. Connections between RLS, DPPs and Christoffel functions were explored in (Fanuel et al., 2019). Note that guarantees for DPP sampling for coresets have been derived in (Tremblay et al., 2019).

2. Main results

2.1. DPP and implicit regularization

Theorem 1 can be found in (Fanuel et al., 2020) and (Mutný et al., 2020) in the context of kernel methods and stochastic optimization respectively. It relates the average of pseudo-inverse of kernel submatrices to a regularization inverse of the full kernel matrix.

Theorem 1 (Implicit regularization)**.** *Let $\mathcal{C}$ be a subset sampled according to $DPP(K/\alpha)$ with $K \succ 0$. Then, we have the identity $\mathbb{E}_C[C K_{\mathcal{CC}}^{-1} C^\top] = (K + \alpha \mathbb{I})^{-1}$.*

Interestingly, a large regularization parmeter $\alpha > 0$ corresponds to small expected subset size $\mathbb{E}[|\mathcal{C}|] = \operatorname{Tr}\left(K(K + \alpha \mathbb{I})^{-1}\right)$. We now discuss an analogous result in the case of kDPPs, for which the implicit regularization effect can be observed.

2.2. Analogous result for kDPP sampling

The elementary symmetric polynomial $e_k(K)$ is proportional to the $(n-k)$-th coefficients of the characteristic polynomial $\det(t\mathbb{I} - K) = \sum_{k=0}^{n} (-1)^k e_k(K) t^{n-k}$. Those polynomials are defined on the vector $\boldsymbol{\lambda}$ of eigenvalues of K. There are explicitly given by the formula $e_k(\boldsymbol{\lambda}) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} \lambda_{i_1} \ldots \lambda_{i_k}$. The kDPPs(K) are defined by the subset probabilities $\Pr(Y = \mathcal{C}) = \det(K_{\mathcal{CC}})/e_k(K)$, and corresponds to DPPs conditioned to a fixed subset size k. Now, we state a result analogous to Theorem 1.

Lemma 1. *Let $\mathcal{C} \sim kDPP(K)$ and $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^n$. We have the identities*

$$\mathbb{E}_C[\boldsymbol{u}^\top C K_{\mathcal{CC}}^{-1} C^\top \boldsymbol{w}] = \frac{e_k(K) - e_k(K - \boldsymbol{w}\boldsymbol{u}^\top)}{e_k(K)}$$
$$= \frac{(-1)^{k+1}}{(n-k)!} \frac{\mathrm{d}^{(n-k)}}{\mathrm{d}\, t^{n-k}} \left[\frac{\boldsymbol{u}^\top \operatorname{adj}(t\mathbb{I} - K) \boldsymbol{w}}{e_k(K)} \right]_{t=0},$$

where adj *is the adjugate of a matrix.*

Proof. Firstly, we use the matrix determinant lemma:

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

By taking the expectation over $\mathcal{C} \sim \text{kDPP}(K)$, we find

$$\mathbb{E}[\boldsymbol{u}^\top C K_{\mathcal{CC}}^{-1} C^\top \boldsymbol{w}] = \frac{e_k(K) - e_k(K - \boldsymbol{w}\boldsymbol{u}^\top)}{e_k(K)} =: \mathcal{E}.$$

where we used that $\sum_{|\mathcal{C}|=k} \det A_{\mathcal{CC}} = e_k(A)$ for any square matrix A. Next, we use the identity $e_k(K) = \frac{(-1)^k}{(n-k)!} \frac{\mathrm{d}^{(n-k)}}{\mathrm{d}\, t^{n-k}} [\det(t\mathbb{I} - K)]_{t=0}$ to obtain the corresponding coefficient of the polynomial $\det(t\mathbb{I} - K) = \sum_{k=0}^n (-1)^k e_k(K) t^{n-k}$. Then, we use once more the matrix determinant lemma with the matrix $(t\mathbb{I} - K)$ this time. This gives

$$\mathcal{E} = \frac{(-1)^{k-1}}{(n-k)!} \frac{\mathrm{d}^{(n-k)}}{\mathrm{d}\, t^{n-k}} \left[\frac{\boldsymbol{u}^\top (t\mathbb{I} - K)^{-1} \boldsymbol{w}}{e_k(K)} \det(t\mathbb{I} - K) \right]_{t=0}.$$

Finally, we recall that $\mathrm{adj}(A) = \det(A) A^{-1}$, which completes the proof. $\square$

The implicit regularization due to the diverse sampling is not explicit in Lemma 1. In order to clarify this formula, we write first an equivalent expression for it. Let the eigendecomposition of K be $K = \sum_{\ell=1}^n \lambda_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top$. Denote by $\boldsymbol{\lambda} \in \mathbb{R}^n$ the vector containing the eigenvalues of K, sorted such that $\lambda_1 \geq \cdots \geq \lambda_n$. Let $\boldsymbol{\lambda}_{\hat{k}} \in \mathbb{R}^{n-1}$ be the same vector with λ_k missing.

Corollary 1. *Let $\mathcal{C} \sim kDPP(K)$. We have the identity:*

$$\mathbb{E}_C[C K_{\mathcal{CC}}^{-1} C^\top] = \sum_{\ell=1}^n \frac{\boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top}{\lambda_\ell + \frac{e_k(\boldsymbol{\lambda}_{\hat{\ell}})}{e_{k-1}(\boldsymbol{\lambda}_{\hat{\ell}})}}. \tag{3}$$

Proof. To begin with, we expand the adjugate in Lemma 1 in the basis of eigenvectors of K. This gives

$$\mathrm{adj}(t\mathbb{I} - K) = \sum_{\ell=1}^n \frac{\prod_{\ell'=1}^n (t - \lambda_{\ell'})}{t - \lambda_\ell} \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top$$

Then, by the definition of the polynomials e_k and by noting that $n - k = n - 1 - (k-1)$, we find

$$\frac{(-1)^{k-1}}{(n-k)!} \frac{\mathrm{d}^{(n-k)}}{\mathrm{d}\, t^{n-k}} \left[\prod_{\ell' \neq \ell} (t - \lambda_{\ell'}) \right]_{t=0} = e_{k-1}(\boldsymbol{\lambda}_{\hat{\ell}}),$$

where $\boldsymbol{\lambda}_{\hat{\ell}} \in \mathbb{R}^{n-1}$ is the vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ with λ_ℓ missing. This yields $\mathbb{E}_C[C K_{\mathcal{CC}}^{-1} C^\top] = \sum_{\ell=1}^n \frac{e_{k-1}(\boldsymbol{\lambda}_{\hat{\ell}})}{e_k(\boldsymbol{\lambda})} \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top$. The final identity is obtained by using the following recurrence relation $e_k(\boldsymbol{\lambda}) = \lambda_\ell e_{k-1}(\boldsymbol{\lambda}_{\hat{\ell}}) + e_k(\boldsymbol{\lambda}_{\hat{\ell}})$. $\square$

It is now possible to illustrate the connection between Corollary 1 and implicit regularization. We give a lower bound for the identity in Corollary 1.

Proposition 1. *With the notations defined above, we have*

$$\mathbb{E}_C[C K_{\mathcal{CC}}^{-1} C^\top] \succeq \sum_{\ell=1}^n \frac{\boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top}{\lambda_\ell + \alpha}, \tag{4}$$

where $\alpha = \sum_{i=k}^n \lambda_i$ and $\mathcal{C} \sim kDPP(K)$.

The above bound matches the expectation formula for DPPs for this specific α. Also, notice that it was remarked in (Dereziński et al., 2020) that if $\alpha = \sum_{i=k}^n \lambda_i$ then $\mathbb{E}_{\mathcal{C} \sim DPP(K/\alpha)}[|\mathcal{C}|] \leq k$. The inequality (4) is obtained thanks to the following Lemma with $l = k$.

Lemma 2 (Eqn 1.3 in (Guruswami & Sinop, 2012))**.** *Let $\boldsymbol{\sigma} \in \mathbb{R}^n$ be a vector with entries $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Let k and l be integers such that $k \geq l > 0$. Then, we have $\frac{e_{k+1}(\boldsymbol{\sigma})}{e_k(\boldsymbol{\sigma})} \leq \frac{1}{k-l+1} \sum_{i=l+1}^n \sigma_i$.*

With the help of Lemma 2, we can prove (4).

Proof of Proposition 1. Let $k \geq 1$. We can lower bound the ratio $\frac{e_{k-1}(\boldsymbol{\lambda}_{\hat{\ell}})}{e_k(\boldsymbol{\lambda}_{\hat{\ell}})}$ in (3) by using Lemma 2. Namely let $\boldsymbol{\sigma}$ be the vector $\boldsymbol{\lambda}_{\hat{\ell}} \in \mathbb{R}^{n-1}$ with entries sorted in decreasing order, and let $l = k$. Then, it holds that $\frac{e_k(\boldsymbol{\sigma})}{e_{k-1}(\boldsymbol{\sigma})} \leq \sum_{i=k}^{n-1} \sigma_i$. By using the definition of $\boldsymbol{\sigma}$, we find that, if $k < \ell$, we have $\sum_{i=k}^{n-1} \sigma_i = -\lambda_\ell + \sum_{i=k}^n \lambda_i$. Otherwise, if $k \geq \ell$, we have $\sum_{i=k}^{n-1} \sigma_i = \sum_{i=k+1}^n \lambda_i$. Hence, we find the upper bound

$$\frac{e_k(\boldsymbol{\sigma})}{e_{k-1}(\boldsymbol{\sigma})} \leq \sum_{i=k}^{n-1} \sigma_i \leq \sum_{i=k}^n \lambda_i = \alpha,$$

since $\boldsymbol{\lambda} \geq 0$. Finally, the statement is proved by using the latter inequality and the identity (3). $\square$

Remark 1 (Upper bound)**.** *Consider the term $\ell = n$ in (3). Then, the additional term at the denominator can be lower bounded as follows:*

$$\frac{e_k(\boldsymbol{\lambda}_{\hat{n}})}{e_{k-1}(\boldsymbol{\lambda}_{\hat{n}})} \geq \frac{n-k}{k} \lambda_{n-1} \left(\frac{\lambda_{n-1}}{\lambda_1} \right)^{k-1} \geq 0,$$

where we used that $e_k(\boldsymbol{\lambda}_{\hat{n}})$ includes $\binom{n-1}{k}$ terms. This bound is pessimistic although it instructs that a small k benefits to the regularization.

As we have observed, the formulae of Theorem 1 or Corollary 1 show that the expectation over diverse subsets implicitly regularize the inverse of the kernel matrix. The improvement of this bound is worth further investigation. A related work (Mutný et al., 2020) uses the same formula given in Theorem 1 to study the convergence of a random block coordinate optimization method for Kernel Ridge Regression, but does not study the ridgeless limit.

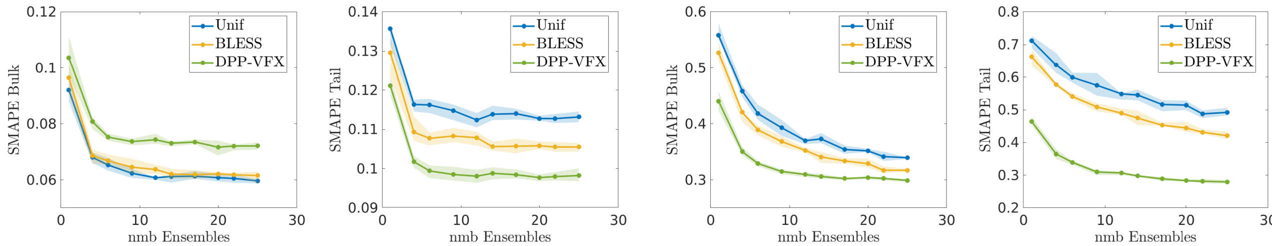

Figure 2. Ensemble KRR on the Bikesharing and CASP dataset (from left to right). The SMAPE on the bulk and tail of the dataset is given in function of the number of ensembles.

3. Experimental results

Sampling a more diverse subset improves the performance of Nyström approximation and KRR (Fanuel et al., 2020). In these experiments, we discuss ensemble approaches for the ridgeless case. The following datasets[1] are used: `Adult`, `Abalone`, `Wine Quality`, `Bike Sharing` and `CASP`. We use 3 sampling algorithms with increasing diversity: uniform sampling, exact ridge leverage score sampling (RLS) (El Alaoui & Mahoney, 2015) and kDPP sampling (Kulesza & Taskar, 2011). For larger datasets the BLESS algorithm (Rudi et al., 2018) is used instead of RLS and DPP-VFX(Dereziński et al., 2019) to speed up the sampling of a kDPP. These algorithms have a relativity small re-sampling cost that motivates their use for ensemble approaches. RLS can be seen as a cheaper proxy for DPP sampling as done in (Dereziński et al., 2020). The different parameters and sample sizes are given in the Supplementary Material. 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 0.25 and 0.75 quantile.

Ensemble Nyström. The accuracy of the approximation is evaluated by calculating $\|K - \hat{K}\|_F/\|K\|_F$ with the ensemble Nyström approximation $\hat{K} = \frac{1}{m}\sum_{i=1}^m KC_i(K_{\mathcal{C}_i\mathcal{C}_i} + \varepsilon\mathbb{I}_k)^{-1}C_i^\top K$ with $\varepsilon = 10^{-12}$ for numerical stability. We illustrate the use of diverse ensembles on Figure 3. Averaging multiple Nyström approximations improves the accuracy. The gain is the most apparent for the more diverse sampling algorithms. Similarly to the experiments in (Kumar et al., 2009), we see that uniform sampling combined with equal mixture weights does not improve performance. This is not the case when using more sophisticated sampling algorithms.

Ensemble KRR. Following the implicit regularization of DPP samplings, we asses the performance of averaging ridgeless predictors trained on DPP subsets. Prediction is done by averaging the ridgeless predictors in an ensemble

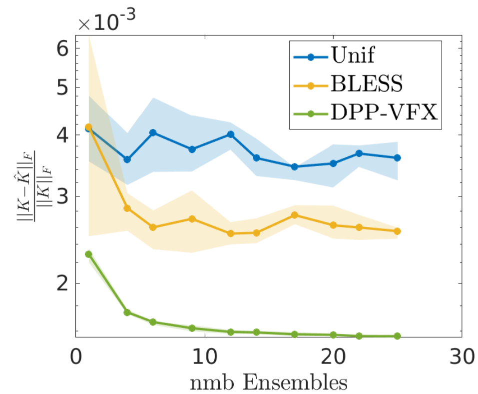

Figure 3. Ensemble Nyström approximation on the Adult dataset. The relative Forbenius norm of the approximation is given in function of the number of ensembles.

approach: $\bar{f} = \frac{1}{m}\sum_{i=1}^m f^\star_{\mathcal{C}_i}$. We evaluate by the same procedure as in (Fanuel et al., 2020). 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. 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 with regularization $\alpha = 10^{-4} \times n_{\text{train}}$ 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. This stratification of the dataset allows to visualize how the regressor performs in dense (small RLS) and sparser (large RLS) groups of the dataset. We calculate the symmetric mean absolute percentage error (SMAPE): $\frac{1}{n}\sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$ of each group. The results for exact sampling algorithms are visualised on Figure 1, approximate algorithms are given on Figure 2. Combining multiple subsets shows a reduction in error. Following (Fanuel et al., 2020), sampling a more diverse subset improves the performance of the KRR. Particularly diverse sampling has comparable performance for the bulk data, while performing much better in the tail of the data. Importantly, all the methods reach a stable performance before the number of points used by all interpolators exceeds the total number of training points.

[1] https://archive.ics.uci.edu/ml/index.php

References

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

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Dereziński, M. and Mahoney, M. W. Determinantal point processes in randomized numerical linear algebra. *arXiv preprint arXiv:2005.03185*, 2020.

Dereziński, M., Liang, F., and Mahoney, M. W. Exact expressions for double descent and implicit regularization via surrogate random design. *arXiv preprint arXiv:1912.04533*, 2019.

Dereziński, M., Khanna, R., and Mahoney, M. W. Improved guarantees and a multiple-descent curve for the Column Subset Selection Problem and the Nyström method. *arXiv preprint arXiv:2002.09073*, 2020.

Desolneux, A., Launay, C., and Galerne, B. Exact sampling of determinantal point processes without eigendecomposition. *arXiv preprint arXiv:1802.08429*, 2018.

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

Fanuel, M., Schreurs, J., and Suykens, J. A. Nyström landmark sampling and regularized Christoffel functions. *arXiv preprint arXiv:1905.12346*, 2019.

Fanuel, M., Schreurs, J., and Suykens, J. A. Diversity sampling is an implicit regularization for kernel methods. *arXiv preprint arXiv:2002.08616*, 2020.

Guruswami, V. and Sinop, A. K. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pp. 1207–1214, 2012.

Hsieh, C.-J., Si, S., and Dhillon, I. A divide-and-conquer solver for kernel support vector machines. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 566–574, 2014.

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

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

Kumar, S., Mohri, M., and Talwalkar, A. Ensemble nystrom method. In *Advances in Neural Information Processing Systems*, pp. 1060–1068, 2009.

Li, C., Jegelka, S., and Sra, S. Efficient sampling for k-determinantal point processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1328–1337, 2016.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Musco, C. and Musco, C. Recursive sampling for the Nyström method. In *Advances in Neural Information Processing Systems 30*, pp. 3833–3845, 2017.

Mutný, M., Dereziński, M., and Krause, A. Convergence analysis of block coordinate algorithms with determinantal sampling. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Poulson, J. High-performance sampling of generic determinantal point processes. *Phil. Trans. R. Soc. A.*, 378: 20190059, 2020.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pp. 5672–5682, 2018.

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

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression. In *Conference on learning theory (COLT)*, pp. 592–617, 2013.

A. Parameters and dataset descriptions

The parameters and datasets used in the simulations can be found in Table 1. The dataset dimensions are given by n and d, σ is the bandwidth of the Gaussian kernel, k the size of the subset. The regularization parameter of the RLS is equal to λ_{RLS}. The parameters for DPP-VFX correspond to $\bar{\mathrm{q}}_{\mathrm{xdpp}}$ and $\bar{\mathrm{q}}_{\mathrm{bless}}$. These are the oversampling parameters for internal Nyström approximation of BLESS and DPP-VFX used to guarantee that everything terminates. Tuning parameters of the BLESS algorithm are q_0, c_0, c_1, c_2.

Table 1. Datasets and parameters used in the experiments.

Dataset	n	d	σ	k	λ_{RLS}	$\bar{\mathrm{q}}_{\mathrm{xdpp}}$	$\bar{\mathrm{q}}_{\mathrm{bless}}$	q_0	c_0	c_1	c_2
`Adult`	48842	110	5	250	10^{-3}	3	3	2	2	3	3
`Abalone`	4177	8	3	50	10^{-4}	/	/	/	/	/	/
`Wine Quality`	6497	11	5	100	10^{-4}	/	/	/	/	/	/
`Bike Sharing`	17389	16	3	250	10^{-3}	3	3	2	2	3	3
`CASP`	45730	9	2	250	10^{-3}	3	3	2	2	3	3