Ensemble Kernel Methods, Implicit Regularization and Determinantal Point Processes (1)

By Joachim Schreurs, Michaël Fanuel and Johan A.K. Suykens (KU Leuven, ESAT-Stadius)

Abstract

Sampling subsets with kDPPs results in implicit regularization in the context of ridgeless Kernel Regression.

Kernel methods

Let $k(x,y) > 0$ be a continuous and *strictly* positive definite kernel. Gram matrix $K = [k(x_i, x_j)]_{i,j}$.
Landmark sampling: $\mathcal{C} \subseteq [n]$.
Sampling matrix: $C \in \mathbb{R}^{n\times|\mathcal{C}|}$, $C = \mathbb{I}_{\mathcal{C}}$.
Submatrices: $K_{\mathcal{C}} = KC$ and $K_{\mathcal{CC}} = C^\top KC$.
Nyström approximation: $L(K,\mathcal{C}) = K_{\mathcal{C}} K_{\mathcal{CC}}^{-1} K_{\mathcal{C}}^\top$.

DPP

Let L be a $n \times n$ positive definite symmetric matrix. The probability of sampling a subset $\mathcal{C} \subseteq [n]$ is

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

- Define $L = K/\alpha$ with $\alpha > 0$.
- denote the process by $DPP_L(K/\alpha)$.

The inclusion probabilities are given by

$$\Pr(\mathcal{C} \subseteq Y) = \det(P_{\mathcal{CC}}),$$

where the marginal kernel is $P = K(K + \alpha\mathbb{I})^{-1}$,. The diagonal of P gives the **Ridge Leverage Scores** (RLS) of the data points: $\boldsymbol{\ell}_i = P_{ii}$ for $i \in [n]$. See El Alaoui, Mahoney, NeurIPS 2015.

Ridgeless regression:

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

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

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 $\boldsymbol{k}_x = [k(x,x_1), \ldots, k(x,x_n)]^\top$.

Implicit regularization with DPP sampling

For $\mathcal{C} \sim DPP(K/\alpha)$, the expectation of the rigdeless predictors gives the function

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

which is the solution of Kernel Ridge Regression

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

A large $\alpha > 0$ yields a small expected subset size for $DPP(K/\alpha)$.

Theorem 1. *Let $\mathcal{C} \sim DPP(K/\alpha)$ with $K \succ 0$. Then,*

$$\mathbb{E}_C[CK_{\mathcal{CC}}^{-1}C^\top] = (K + \alpha\mathbb{I})^{-1}.$$

Notice that $(CK_{\mathcal{CC}}C^\top)^\dagger = CK_{\mathcal{CC}}^{-1}C^\top$.
(see, Fanuel, Schreurs, Suykens arXiv:1905.12346 and Mutný, Dereziński, Krause AISTATS 2020)

Analogous results for kDPPs

kDPPs(K) are defined by

$$\Pr(Y = \mathcal{C}) = \det(K_{\mathcal{CC}})/e_k(K),$$

where $e_k(\boldsymbol{\lambda}) = \sum_{1\leq i_1 < \cdots < i_k \leq n} \lambda_{i_1} \ldots \lambda_{i_k}$ are elementary symmetric polynomials.

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

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

The above result is easier to interpret in the spectral domain.

Understanding Lemma 1

Let the eigendecomposition of K be

$$K = \sum_{\ell=1}^n \lambda_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top.$$

Denote by $\boldsymbol{\lambda} \in \mathbb{R}^n$ contain the eigenvalues of K, 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:*

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

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

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

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

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

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

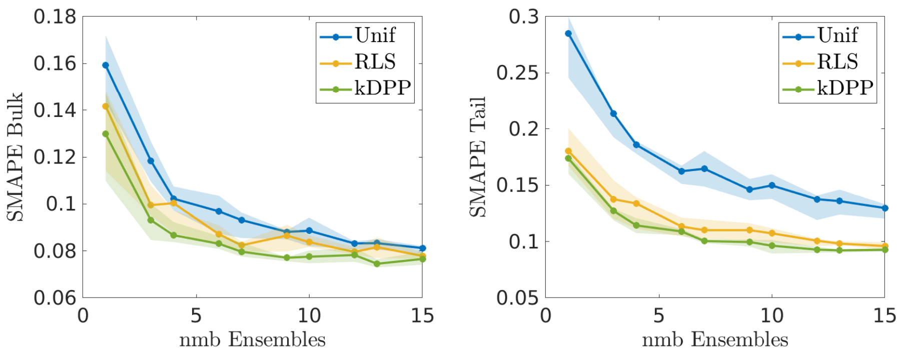
where we used that $e_k(\boldsymbol{\lambda}_{\hat{n}})$ includes $\binom{n-1}{k}$ terms.

European Research Council
Established by the European Commission

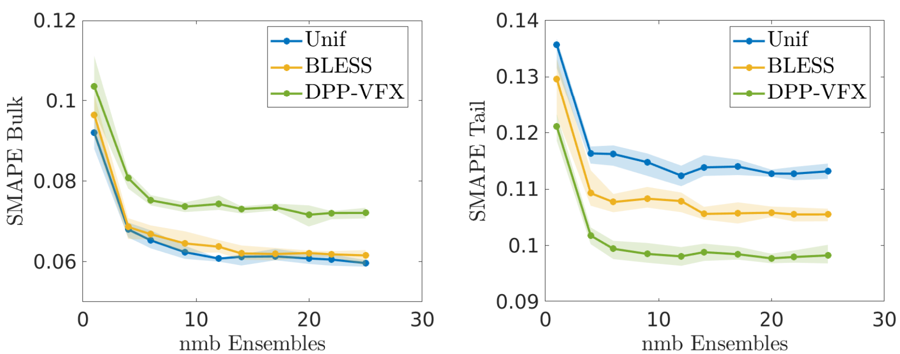
Ensemble Kernel Methods, Implicit Regularization and Determinantal Point Processes (2)

By Joachim Schreurs, Michaël Fanuel and Johan A.K. Suykens (KU Leuven, ESAT-Stadius)

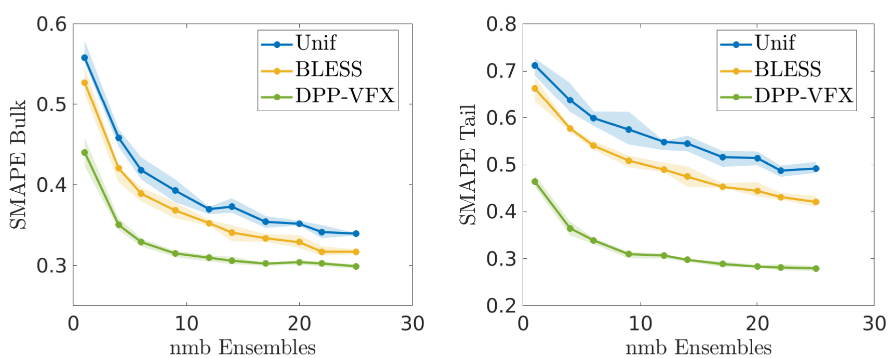
Ensemble Ridgeless Regressions: Abalone dataset



Ensemble Ridgeless Regressions: Bikesharing dataset



Ensemble Ridgeless Regressions: CASP dataset



Methodology

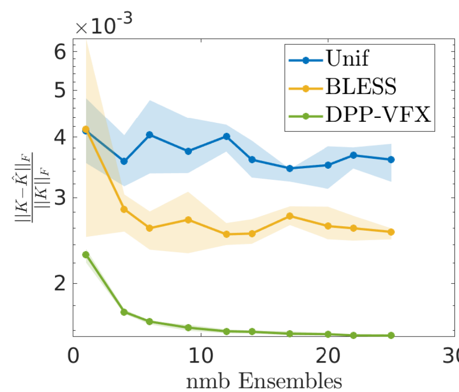
- Prediction is done by averaging the ridgeless predictors in an ensemble approach: $\bar{f} = \frac{1}{m}\sum_{i=1}^{m} f^{\star}_{\mathcal{C}_i}$.
- Split in 50% training and 50% test data.
- The dataset is stratified: the test set is divided into 'bulk' and 'tail'.
  - Bulk: test points where RLS are smaller than the 70% quantile
  - Tail: test points where RLS are larger than the 70% quantile.
- 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.

Comparisons of different samplings

We use 3 sampling algorithms:

- uniform sampling,
- exact RLS sampling and approximate RLS: BLESS (Rudi et al. NeurIPS 2018).
- exact kDPP sampling and approximate kDPP: DPP-VFX (Derezinski et al. NeurIPS 2019).

Ensemble Nyström: Adult dataset



Datasets

Dataset	n	d
Adult	48842	110
Abalone	4177	8
Wine Q.	6497	11
Bike S.	17389	16
CASP	45730	9

Conclusions

Exact formula for implicit regularization. Interest for regression problems in order to achieve a low MAPE in the tail of the test data.

Ensemble Kernel Methods, Implicit Regularization and Determinantal Point Processes (3)

By Joachim Schreurs, Michaël Fanuel and Johan A.K. Suykens (KU Leuven, ESAT-Stadius)

References

- Alaoui, A. and Mahoney, M. W., Fast randomized kernel ridge regression with statistical guarantees. Neurips 2015.
- Bach F., Sharp analysis of low-rank kernel matrix approximations, COLT 2013.
- Belkin, M., Rakhlin, A. and Tsybakov, A.B., Does data interpolation contradict statistical optimality? PMLR 2019.
- Dereziński, M., Calandriello, D., and Valko, M., Exact sampling of determinantal point processes with sublinear time preprocessing, NeurIPS 2019.
- Fanuel, M., Schreurs, J., and Suykens, J. A., Diversity sampling is an implicit regularization for kernel methods. arXiv:2002.08616.
- Fanuel, M., Schreurs, J. and Suykens, J., Nyström landmark sampling and regularized Christoffel functions. arXiv:1905.12346.
- Kulesza, A. and Taskar, B., Determinantal point processes for machine learning. Foundations and Trends in Machine Learning 5.2–3: 123-286, 2012.
- Li, C., Jegelka, S. and Sra, S. , Fast DPP sampling for Nyström with application to kernel methods, ICML 2016.
- Liang, T., and Rakhlin, A., Just interpolate: Kernel "ridgeless" regression can generalize. Annals of Statistics, arXiv:1808.00387.
- Mutný, M., Dereziński, M., and Krause, A. Convergence analysis of block coordinate algorithms with determinantal sampling, AISTATS 2020.
- Rudi, A., Calandriello, D., Carratino, L. and Rosasco, L., On Fast Leverage Score Sampling and Optimal Learning, NeurIPS 2018.

Acknowledgments

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. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. 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.