Determinantal Point Process for Nyström approximation of Gaussian Process Regression

Kei Ishikawa KEII@STUDENT.ETHZ.CH

Abstract

We investigate the application of Nyström method and determinantal point process on Gaussian process regression. We theoretically show that uncertainty quantification performance of approximated Gaussian process measured by in-sample KL-divergence from the full Gaussian process can be bounded by the spectrum norm error of Nyström approximation. In the numerical experiment, we show that the combination of Nyström method and determinantal point process gives the best approximation of Gaussian process regression under certain situations.

Keywords: determinantal point process, Nyström approximation, Gaussian process regression

1. Introduction

Kernel machines have been successfully applied in biology and medicine due to well-known similarity metrics on discrete structures, such as DNA sequence and molecular compounds. In most of the successful applications, support vector machines (SVM) are adopted because of their good performance. However, SVMs do not typically provide uncertainty quantification which is a crucial feature for applications such as experimental design. Gaussian processes (Rasmussen and Williams (2006)) can be thought of as an alternate type of kernel machine that provides better uncertainty estimates. There are efficient algorithms for scaling Gaussian process that leverages the gradient-based optimization. However, Gaussian process on discrete domains has still been limited by their cubic time-complexity at prediction time because of the non-differentiability of the kernel. So the main motivation for this research is to establish a method for scaling Gaussian process in such setting where kernel is not differentiable.

There are mainly two approaches to scaling the Gaussian process. Both approaches rely on inducing points set chosen from the data domain and offer linear running time in data size. Roughly speaking, inducing points set is used to summarize the information of the whole dataset on it, so its size is usually chosen much smaller than that of the original dataset and it needs to represent the whole dataset well. One approach for scaling the Gaussian process is sparse variational approximation (Titsias (2009a,b)). It makes an assumption that the posterior distribution can be expressed as a conditional GP given densities at inducing points and then use variational inference obtain approximate posterior. The sparse variational approximation is particularly powerful when differentiable kernel such as RBF kernel is used because inducing points can be chosen using gradient-based optimization technique. Hensman et al. (2013) extends this direction by combining stochastic optimization. The other approach is Nyström method(Williams and Seeger (2001)). This method has wide applicability for scaling many kinds of kernel machines as it just makes the operations on kernel faster by low-rank matrix approximation. In the Nyström method, all data points are projected to the subspace spanned by inducing points in the feature space. This result in a great reduction of complexity for matrix operations such as matrix inversion. The detail of the complexity of the Nyström method is discussed in the section 2.

As can be expected, the choice of the inducing points is very important for the performance of the approximated Gaussian process. There is a lot of work about the inducing point selection for Nyström method (Mahoney et al. (2011); Drineas et al. (2012); Gittens and Mahoney (2016)). Recently, a new sampling technique that uses determinantal point process(DPP) has been proposed for inducing points selection (Belabbas and Wolfe (2009a,b)). (We call this technique DPP-Nyström for simplicity in this paper.) As discussed later, DPP gives strong statistical guarantees on the trace norm of the approximation error. It also allows for linear complexity in sampling (Anari et al. (2016); Li et al. (2016a)).

There are only a small number of research that applies the determinantal point process to Gaussian process. Very recently, Burt et al. (2019, 2018) have shown the upper bound on KL-divergence between an approximated in-sample posterior distribution and true posterior (posterior of full Gaussian process), in the case of sparse variational approximation. In the context of the application of Nyström method to other kernel machines, there are several works about the in-sample error analysis for kernel ridge regression (Bach (2013); Alaoui and Mahoney (2015); Li et al. (2016a)). However, as far as we know, there is no research that combines DPP-Nyström method and Gaussian process regression.

In this work, we study the uncertainty quantification performance of Gaussian process approximated by DPP-Nyström method and show that DPP-Nyström method is suited for large scale approximation of Gaussian process. In the later section, we empirically show that DPP-Nyström method is a scheme for uncertainty quantification of Gaussian process. We also show the upper bounds for the in-sample KL divergence between a full model and a low-rank model.

In section 2, we first review Nyström method, determinantal point process and some known properties about DPP-Nyström. We also review notations and the definitions of the Gaussian process. In section 4, we present our analysis on the approximation performance of DPP-Nyström for Gaussian process regression. In section 5, we empirically show that DPP-Nyström method performs well for uncertainty quantification.

2. Background and Notation

Throughout this report, we consider the approximation of positive semidefinite kernel matrix $K \in \mathbb{R}^{n \times n}$. We assume that K has an Cholesky decomposition, $K = B^T B$ and $B_{\cdot,I}$ stands for a submatrix of B, corresponding to the column vectors of B whose columns is index is included in I. In addition, we will use b_i for B's i-th column vector. In this notation, the vector b_i can be understood can be viewed as the feature vector of the i-th sample, since the element of kernel matrix $K_{i,j}$ is inner product of b_i and b_j. By $K_{i,\cdot}$ and $K_{\cdot,j}$, we indicate the i-th row vector and j-th colums vector of K. Similarly, given index set $I = \{i_1, ..., i_k\} \in [n]$ of size k, we use $K_{I,\cdot} \in \mathbb{R}^{k \times n}$ and $K_{\cdot,I} \in \mathbb{R}^{n \times k}$ to express submatrix of K whose row/column indices corresponds to I. Moreover, $K_I \in \mathbb{R}^{k \times k}$ denotes a submatrix of K whose row/column indices corresponds to I. Lastly, we use $||\cdot||$, $||\cdot||_F$ for spectral norm and Frobenius norm of matrix.

2.1 Nyström approximation

Nyström approximation is a low rank approximation of positive semidefinite matrix. For given inducing point set $I = \{i_1, ..., i_k\}$ the Nyström approximation of the K can be written as

$$\tilde{K} = K_{\cdot,I} K_I^{-1} K_{I,\cdot} \tag{1}$$

It is clear that by this approximation, we obtain a rank m approximation of matrix K. It now remains to see why this is a good approximation of the original matrix K. As $K_{I,\cdot} = K_{\cdot,I}{}^T = B_{\cdot,I}{}^T B$ and $K_I = B_{\cdot,I}{}^T B_{\cdot,I}$, we have

$$\begin{aligned}\tilde{K} &= B^T B_{\cdot,I}(B_{\cdot,I}{}^T B_{\cdot,I})^{-1} B_{\cdot,I}{}^T B \\ &= B^T P_I B \\ &= (P_I B)^T (P_I B)\end{aligned} \tag{2}$$

where $P_I = B_{\cdot,I}(B_{\cdot,I}{}^T B_{\cdot,I})^{-1} B_{\cdot,I}{}^T$. We can see that P_I is a definition of projector to the linear subspace spanned by column vectors of $B_{\cdot,I}$. Thus, we see that Nyström approximation can be considered as a projection of feature vector to low dimentional linear subspace. Actually, we see that $K_{i,j} = b_i{}^T b_j$ and $\tilde{K}_{i,j} = (P_I b_i)^T (P_I b_j)$. Indeed, this argument can be fully applied to the feature space of the kernel and it can be understood as a special case of kernel principal component analysis (Schölkopf et al. (1997)).

It is known that best k rank approximation in terms of Frobenius norm and spectral norm can be obtained by truncated eigenvalue decomposition of kernel matrix up to m rank. However, in order to compute the truncated eigenvalue decomposition up to rank k, we need $\mathcal{O}(n^2 k)$ time complexity, which is much more prohibitive in the large scale applications compared to linear complexity algorithms.

2.2 Determinantal Point Process

A determinantal point process (DPP) is a probability model over the subset of indices. The probability of process taking value $I = \{i_1, ..., i_k\} \in [n]$ is proportional to the determinant of K_I, which is $\mathrm{P(I) \propto det(K_I)}$. When we fix the cardinality of the subset as $|I| = k$, we get conditional DPP which is called k-DPP. DPP can be said to have a negative correlation between arbitrary two indices in a sense that similar indices are less likely to co-occur.

k-DPP needs $\mathcal{O}(n^3)$ complexity for sampling. However, there is also a approximate Gibbs sampling algorithm with $\mathcal{O}(k^2)$ complexity per iteration. Li et al. (2016b) proposed faster algorithm for the iteration of Gibbs sampler of k-DPP. This enables efficient sampling of k-DPP for inducing points selection. Anari et al. (2016) has shown that mixing time of Gibbs sampling algorithm for k-DPP is $\mathcal{O}(nk)$. Mixing time is a number of iteration in sampling required for the Markov Chain's convergence to the stationary distribution. Thus the total complexity of approximate sampling of k-DPP is $\mathcal{O}(nk^3)$. The detail of algorithm is shown in the Algorithm 1.

```
Algorithm 1 Gibbs sampler for $k$-DPP (Li et al. (2016a))
  Input: $K$ the kernel matrix, $[n] = \{1, ..., n\}$ the ground set
  Output: $I$ sampled from exact $k$-DPP$(K)$
  Randomly Initialize $I \subseteq [n]$, $|I| = k$
  while not mixed do
    Sample $b$ from uniform Bernoulli distribution
    if $b = 1$ then
      Pick $j^{\mathrm{in}} \in I$ and $j^{\mathrm{out}} \in [n] \backslash I$ uniformly randomly
      $q(j^{\mathrm{in}}, j^{\mathrm{out}}, I) \leftarrow \frac{\det(K_{I \cup \{j^{\mathrm{out}}\} \backslash \{j^{\mathrm{in}}\}})}{\det(K_{I \cup \{j^{\mathrm{out}}\} \backslash \{j^{\mathrm{in}}\}}) + \det(K_I)}$
      $I \leftarrow I \cup \{j^{\mathrm{out}}\} \backslash \{j^{\mathrm{in}}\}$ with prob. $q(j^{\mathrm{in}}, j^{\mathrm{out}}, I)$
    end if
  end while
```

In Kulesza et al. (2012) 2.2.1, intuitive interpretation of DPP is given. By using the Cholesky decomposition, $K = B^T B$, we can prove that

$$\det(K_I) = \mathrm{vol}(B_{\cdot,I})^2 \tag{3}$$

$$= \mathrm{vol}(\{b_i | i \in I\})^2 \tag{4}$$

Here, $\mathrm{vol}(\mathrm{B}_{\cdot,\mathrm{I}})$ stands for the volume of parallelepiped spanned by the column vectors of $B_{\cdot,I}$. Thus, we can expect that DPP will sample indices $i_1, ..., i_k$ with diverse feature vectors $b_{i_1}, ..., b_{i_k}$.

2.3 DPP-Nyström Method

We propose to select the inducing point set for the Nyström method by k-DPP. By doing so, we can expect to get diverse inducing points that represents the whole dateset well. In this report, we will call this the DPP-Nyström method. As discussed above, by k-DPP, we can expect to get B_I with more diverse column vectors. Intuitively, it sounds reasonable to combine k-DPP to Nyström approximation as Nyström approximation uses projected column vector $P_I b_1, ..., P_I b_n$ to obtain low-rank approximation.

The use of DPP for the inducing point selection was first introduced by Belabbas and Wolfe (2009b,a). They derived following upper bound on the trace error.

Lemma 1. (Belabbas and Wolfe (2009b,a)) When the inducing point set $I = \{i_1, ..., i_k\}$ is sampled by k-DPP,

$$\mathbb{E}_{I \sim \text{k-DPP}} \left[\mathrm{tr}(K - \tilde{K}) \right] \leq (k+1) \sum_{i=k+1}^{n} \lambda_i \tag{5}$$

From the above upper bound, we see that DPP-Nyström method achieves low trace norm error when k is chosen so that $\lambda_i \simeq 0$ for all $i \geq k + 1k$. In other words, error becomes small when the size of inducing point set is large than effective rank of K, the error becomes fairly small. Recently, Li et al. (2016a) also gave upper bound on Frobenius norm error.

Lemma 2. Li et al. (2016a) When the inducing point set $I = \{i_1, ..., i_k\}$ is sampled by k-DPP, for all $l < k$,

$$\mathbb{E}_{I\sim\text{k-DPP}}\left[||K-\tilde{K}||_F\right] \leq \frac{k+1}{k+1-l}\sum_{i=l+1}^{n}\lambda_i \tag{6}$$

This bound also implies that when k is chosen to be bigger than effective rank of K, the Frobenius norm error becomes small.

2.4 Gaussian Process Regression

Gaussian process regression is a kernel machine which provides reasonable uncertainty estimate of the prediction. In the Gaussian process regression, we assume that the $y_1, ..., y_m$ are noisy observations of $f_1, ..., f_m$. The observational noise is mean-zero Gaussian additive noise. $f_1, ..., f_m$ are assumed to be generated by a Gaussian process whose mean is zero and the covariance matrix is K. Here, $K_{i,j} = K(x_i, x_j)$ and $K(\cdot, \cdot)$ is a kernel. Therefore, we can write for $i = 1, ..., n$,

$$y_i = f_i + \varepsilon_i \tag{7}$$

$$f_{1:n} \sim \mathcal{GP}(0, K) \tag{8}$$

$$\varepsilon_i \sim_{\text{ i.i.d. }} N(0, \tau^2) \tag{9}$$

Then prediction density of the Gaussian process regression can be written as $p(f_{1:n}|y_{1:m}) = N(\mu_{1:n|1:m}, \Sigma_{1:n|1:m})$ where

$$\mu_{1:n|1:m} = K_{1:n,1:m}(K_{1:m,1:m} + \tau^2 I_m)^{-1}y \tag{10}$$

$$\Sigma_{1:n|1:m} = K_{1:n,1:n} - K_{1:n,1:m}(K_{1:m,1:m} + \tau^2 I_m)^{-1}K_{1:m,1:n} \tag{11}$$

Here, $y = (y_1, ..., y_m)^T$. We call prediction for $f_{1:m} = (f_1, ..., f_m)^T$ as in-sample prediction and those for $f_{m+1:n} = (f_{m+1}, ..., f_n)^T$ as out-of-sample prediction.

3. Proposed Method

The DPP-Nyström method can be combined with Gaussian process regression straightforwardly by replacing kernel matrix with the approximated matrix. The prediction density of approximated Gaussian process now becomes $p(f_{1:n}|y_{1:m}) = N(\tilde{\mu}_{1:n|1:m}, \tilde{\Sigma}_{1:n|1:m})$ where

$$\tilde{\mu}_{1:n|1:m} = \tilde{K}_{1:n,1:m}(\tilde{K}_{1:m,1:m} + \tau^2 I_m)^{-1}y \tag{12}$$

$$\tilde{\Sigma}_{1:n|1:m} = K_{1:n,1:n} - \tilde{K}_{1:n,1:m}(\tilde{K}_{1:m,1:m} + \tau^2 I_m)^{-1}\tilde{K}_{1:m,1:n} \tag{13}$$

Here, we are combining the use of $\tilde{K}$ and K for kernel matrix. This helps to reduce the underestimation of predictive variance. It also keeps the predictive covariance matrix positive semidefinite as $K - \tilde{K}$ is always positive semidefinite. By using k-DPP at the inducing points selection, we get an approximation of Gaussian process which has linear complexity and theoretical guarantee. We will later show empirically that this method performs well. In the project, the use of simulated annealing was tried. This is because it intuitively makes sense to use inducing points set whose determinant is as big as possible as determinant can be understood as a measure of diversity of feature vectors of the points set. Nevertheless, simulated annealing turned out to be a little inferior method for inducing points selection than k-DPP from the empirical experiment as discussed in the section 5.

4. Theoretical Analysis

In order to evaluate the uncertainty quantification ability of approximated Gaussian process, we use KL-divergence between approximated posterior over a data point and posterior over data point of full Gaussian process. This quantity is used to evaluate the quality of the approximation of the Gaussian process by Burt et al. (2019, 2018). They derive upper bound on this quantity in the case of sparse variational approximation and we get very similar results for Nyström approximation of Gaussian process. We will first review their results.

Theorem 1. (Burt et al. (2018) Lemma 1.)
Let $p(f|y)$ be a true in-sample posterior of Gaussian process given observation y and let $q(f|y)$ be a approximate in-sample posterior of Gaussian process obtained by sparse variational approximation (Titsias (2009a,b)) using inducing point set I, then for all I,

$$KL\left[q(f|y)||p(f|y)\right] \leq \frac{t}{2\tau^2}\left(1+\frac{||y||_2^2}{\tau^2+t}\right) \tag{14}$$

where $t = \mathrm{tr}(\mathrm{K}-\tilde{\mathrm{K}})$ is the only term which is dependent on I. .

This result indicates that the approximation of Gaussian process gets better as the Nyström approximation using the same inducing points set gets more accurate as t is a trace error of Nyström approximation measured by trace.

In our work, we bound the same quantity in the case Nyström approximation is applied to Gaussian process. Our result indicates that the same quantity can be bounded by the error of kernel matrix measured by the spectral norm. In the following discussion, $||A||$ stands for the spectral norm of matrix A.

Theorem 2.

$$\begin{aligned} &KL\left[N(\tilde{\mu}_{1:m|1:m}, \tilde{\Sigma}_{1:m|1:m})||N(\mu_{1:m|1:m}, \Sigma_{1:m|1:m})\right] \\ &\leq \left(\frac{9}{2}\mathrm{tr}(\Sigma_{1:m|1:m}^{-2}) + \frac{2||y||_2^2}{\tau^4}\cdot||\Sigma_{1:m|1:m}^{-1}||\right)\cdot||\tilde{K}-K||^2 + \mathcal{O}(||\tilde{K}-K||^3) \end{aligned} \tag{15}$$

.

With the second theorem, we can approximately bound the square root of this in-sample KL-divergence in expectation using 5 and 6. In doing so, we use the property that the spectral norm of the matrix is always smaller than the Frobenius norm and trace norm of the matrix. This suggests that the uncertain quantification performance of approximated Gaussian process becomes almost as good as full Gaussian process when we take the size of inducing points set to be bigger than effective rank of the kernel matrix.

In order to show the main result, we use the following bound of $||\tilde{\mu}_{1:m|1:m} - \mu_{1:m|1:m}||_2$ and $||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}||$ in terms of $||\tilde{K}-K||$. This bound should be useful than KL-divegence in cases where explicit deviation of mean and the covariance matrix is needed.

Lemma 3. Using the same notation as above for the parameters of posterior distributions, we have

$$||\tilde{\mu}_{1:m|1:m} - \mu_{1:m|1:m}||_2 \leq \frac{2||y||_2^2}{\tau^2}||\tilde{K}-K|| \tag{16}$$

$$||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}|| \leq 3||\tilde{K} - K|| \tag{17}$$

The proof for theorem 2. and lemma 3. is given in the Appendix 6.

5. Experiments

In the numerical experiment, we compared the performance of Nyström approximation for Gaussian process regression using different inducing point selection schemes. Throughout the experiment, the ailerons dataset (ail) was used, though models and dataset from Fortuin et al. (2018) was originally planed to be used. This change in the dataset is made because the models and dataset used there were mainly for classification. The uncertainty quantification performance of Gaussian process was easier to analyze for regression case. The code for the experiment is available at github.com/kstoneriv3/dpp. We used 1) uniform sampling, 2)k-DPP (by Gibbs sampling), 3)Simulated annealing of DPP, 4)Greedy selection. Here, we did not use kernel k-means clustering (Dhillon et al. (2004)) as a benchmark as kernel k-means clustering requires differentiable kernel. Simulated annealing is based on the Gibbs sampling algorithm of k-DPP and is intended to find the maximum a posteriori (MAP) of k-DPP. The use of the greedy method for inducing points selection of Gaussian process was proposed in the Fortuin et al. (2018). The greedy selection scheme was designed so that the likelihood of the approximated Gaussian process is maximized.

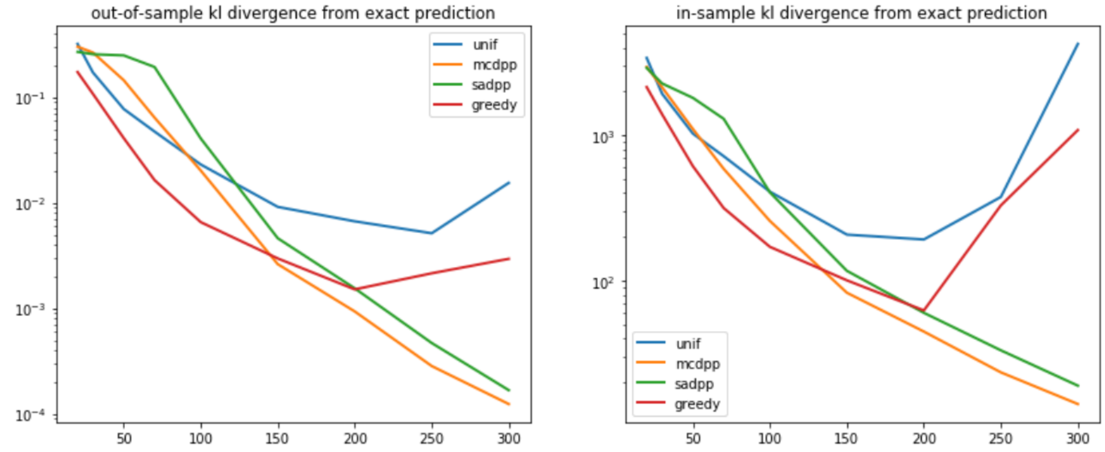

Figure 1: KL-divergence of the prediction density from the prediction of the full Gaussian process.

The figure 1 compares the out-of-sample and in-sample KL-divergence of the predictive density of approximated Gaussian process and that of full Gaussian process. The in-sample KL-divergence is the KL-divergence for in-sample prediction density and the out-of-sample prediction density is that for out-of-sample prediction density. Both out-of-sample and in-sample KL-divergence and in-sample KL show the same tendency. For inducing points set size up to about 150, the greedy algorithm performs the best among all schemes. However, when the size of inducing points set is bigger than around 200, the greedy algorithm and uniform sampling start degrading. This is because these methods give too small variance as a prediction. In other words, those methods become too confident about their prediction when the number of inducing points. In contrast, k-DPP and simulated annealing do not degrade and increase the uncertainty quantification performance as the number of inducing points increases. This suggests that k-DPP is a more promising scheme for inducing points sampling than other benchmarks given here when we can have relatively large number of inducing points. We also observe that simulated annealing for k-DPP does work a bit worse than or almost as good as k-DPP. This implies that simply maximizing the determinant corresponding to inducing points set does not necessarily improve the performance.

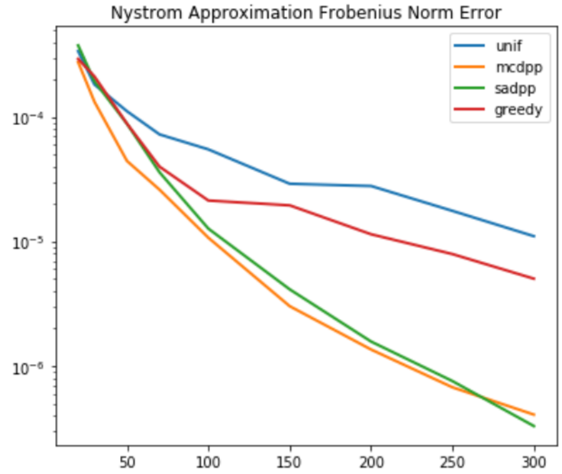

Figure 2: Error of Nyström approximation for each inducing point selection schme. Frobenius norm of the error ($||\tilde{K} - K||_F$) is used.

The figure 2 shows the error of Nyström approximation measured by Frobenius norm. It is clear that k-DPP and simulated annealing of the MAP of the k-DPP works really well for reducing the error. On the other hand, the approximation error of the greedy algorithm and uniform sampling improves slower where the number of inducing points are bigger than 100.

In figure 3, running time of each inducing points selection scheme excepting uniform sampling is presented. The number of iteration for the Gibbs sampler for k-DPP, simulated annealing k-DPP and greedy algorithm is chosen so that each algorithm go over whole sample points once on average. It is clear that the greedy algorithm is much faster than k-DPP or simulated annealing of MAP of k-DPP. When the size of inducing points is about 300, the greedy method is more than 100 times faster than k-DPP and simulated annealing. This is counter-intuitive because the theoretical time complexity for one iteration of the greedy algorithm is $\mathcal{O}(nk^2 + k^3)$ while it is only $\mathcal{O}(k^2)$ for k-DPP and simulated annealing. In our experiment, Gibbs sampling algorithm for k-DPP and simulated annealing algorithm are implemented in a crude Python code while for the greedy algorithm, we mainly used Numpy library, which is a highly optimized linear algebra package for Python. So, it is

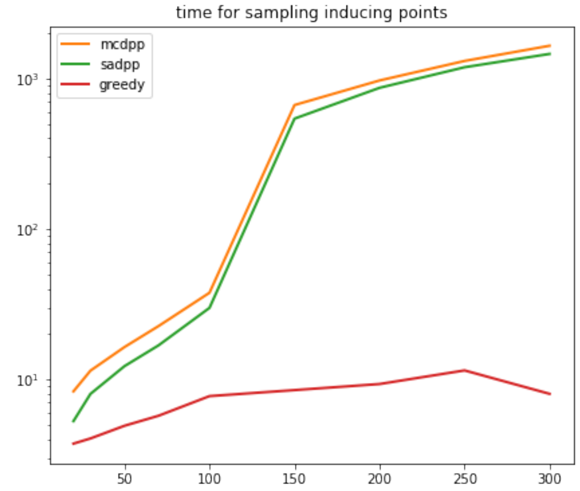

Figure 3: The time used for sampling the inducing points for each sampling scheme. The time cost for uniform sampling is negligible compared to other sampling schemes so it is not included in the plot.

possible that highly optimized computation of Numpy speeded up the greedy algorithm by great extent while crude Python code is very slow for the other two schemes. Another possibility is that constant factor of the complexity of DPP sampler is large. This can be the case because an iterative algorithm is used to compare the determinant at each iteration in the MCMC sampler of k-DPP. In addition, in the numerical experiment, the sample size n was 5000 so that we can compare the KL-divergence with full Gaussian process. Thus, even though the greedy algorithm seems superior in this experiment, k-DPP and simulated annealing might perform faster with efficient implementation and with a bigger sample size n.

6. Conclusion

In this work, we investigated the performance of DPP-Nyström method applied to Gaussian process regression and showed its advantages. We analyzed the performance theoretically and derived upper bound for in-sample KL-divergence, which measures the performance of uncertainty quantification. We also empirically showed that DPP-Nyström method performs better than other benchmarks including greedy algorithm when the number of inducing points is bigger than some threshold. As discussed in the introduction, the final goal of this research is to establish a scheme for scaling Gaussian process with discrete kernel. Thus, applying the DPP-Nyström method to Gaussian process on discrete dataset would be the next step of this research.

Appendix A.

In this appendix we prove the following theorem from section 4. As stated in the section 2, $||\cdot||$ is used for spectral norm.

Lemma 3.

$$
\begin{aligned}
||\tilde{\mu}_{1:m|1:m} - \mu_{1:m|1:m}||_2 &\leq \frac{2||y||_2}{\tau^2}||\tilde{K} - K|| \\
||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}|| &\leq 3||\tilde{K} - K||
\end{aligned}
$$

proof For the notational simplicity, we use $A = K_{1:m,1:m}$ and $\tilde{A} = \tilde{K}_{1:m,1:m}$.

$$
\begin{aligned}
&||\tilde{\mu}_{1:m|1:m} - \mu_{1:m|1:m}||_2 \\
&= ||\tilde{K}_{1:m,1:m}(\tilde{K}_{1:m,1:m} + \tau^2 I_m)^{-1}y - K_{1:m,1:m}(K_{1:m,1:m} + \tau^2 I_m)^{-1}y||_2 \\
&= ||\tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}y - A(A + \tau^2 I_m)^{-1}y||_2 \\
&\leq ||(\tilde{A} - A)(\tilde{A} + \tau^2 I_m)^{-1}y||_2 \\
&\quad + ||A((\tilde{A} + \tau^2 I_m)^{-1} - (A + \tau^2 I_m)^{-1})y||_2 \\
&= ||(\tilde{A} - A)(\tilde{A} + \tau^2 I_m)^{-1}y||_2 \\
&\quad + ||A(A + \tau^2 I_m)^{-1}((A + \tau^2 I_m) - (\tilde{A} + \tau^2 I_m))(\tilde{A} + \tau^2 I_m)^{-1}y||_2 \\
&\leq ||\tilde{A} - A|| \cdot ||(\tilde{A} + \tau^2 I_m)^{-1}|| \cdot ||y||_2 \\
&\quad + ||A(A + \tau^2 I_m)^{-1}|| \cdot ||\tilde{A} - A|| \cdot ||(\tilde{A} + \tau^2 I_m)^{-1}|| \cdot ||y||_2 \\
&\leq ||\tilde{K} - K|| \cdot ||\tau^{-2} I_m|| \cdot ||y||_2 + ||I_m|| \cdot ||\tilde{K} - K|| \cdot ||\tau^{-2} I_m|| \cdot ||y||_2 \\
&= \frac{2||y||_2}{\tau^2}||\tilde{K} - K||
\end{aligned}
$$

and

$$
\begin{aligned}
&||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}|| \\
&= ||(K_{1:m,1:m} - \tilde{K}_{1:m,1:m}(\tilde{K}_{1:m,1:m} + \tau^2 I_m)^{-1}\tilde{K}_{1:m,1:m}) \\
&\qquad - (K_{1:m,1:m} - K_{1:m,1:m}(K_{1:m,1:m} + \tau^2 I_m)^{-1}K_{1:m,1:m})|| \\
&= ||(A - \tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}\tilde{A}) \\
&\qquad - (A - A(A + \tau^2 I_m)^{-1}A)|| \\
&= ||A(A + \tau^2 I_m)^{-1}A - \tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}\tilde{A})|| \\
&\leq ||(A - \tilde{A})(A + \tau^2 I_m)^{-1}A|| \\
&\quad + ||\tilde{A}((A + \tau^2 I_m)^{-1} - (\tilde{A} + \tau^2 I_m)^{-1})A|| \\
&\quad + ||\tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}(A - \tilde{A})|| \\
&\leq ||(A - \tilde{A})(A + \tau^2 I_m)^{-1}A|| \\
&\quad + ||\tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}((\tilde{A} + \tau^2 I_m) - (A + \tau^2 I_m))(A + \tau^2 I_m)^{-1}A|| \\
&\quad + ||\tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}(A - \tilde{A})|| \\
&\leq ||A - \tilde{A}|| \cdot ||(\tilde{A} + \tau^2 I_m)^{-1}\tilde{A}||
\end{aligned}
$$

$$
\begin{aligned}
&+ ||\tilde{A}(\tilde{A} + \tau^2 I_m)^{-1}|| \cdot ||A - A|| \cdot ||(A + \tau^2 I_m)^{-1} A|| \\
&+ ||A(A + \tau^2 I_m)^{-1}|| \cdot ||\tilde{A} - A|| \\
\leq\, & ||K - \tilde{K}|| \cdot ||I_m|| \\
&+ ||I_m|| \cdot ||\tilde{K} - K|| \cdot ||I_m|| \\
&+ ||I_m|| \cdot ||\tilde{K} - K|| \\
=\, & 3||\tilde{K} - K||
\end{aligned}
$$

In the proof, we used the property that $I_m - A(A +^2 I_m)^{-1}$ is positive semidefinite, for all positive semidefinite matrix $A \in \mathbb{R}^{m \times m}$. This can be easily checked using the eigenvalue decomposition of A, which is $A = V \Lambda V^T$, where V is orthonormal matrix and Λ is diagonal matrix with non-negative entry.

Theorem 2.

$$
KL\left[p_{\tilde{K}}(f|y)||p_K(f|y)\right] \leq \left(\frac{9}{2}\mathrm{tr}(\Sigma^{-2}_{1:m|1:m}) + \frac{2||y||_2^2}{\tau^4} \cdot ||\Sigma^{-1}_{1:m|1:m}||\right) \cdot ||\tilde{K} - K||^2 + \mathcal{O}(||\tilde{K} - K||^3)
$$

.

proof

The KL-divergence we consider has following analytical expression.

$$
\begin{aligned}
& KL\left[N(\tilde{\mu}_{1:m|1:m}, \tilde{\Sigma}_{1:m|1:m})||N(\mu_{1:m|1:m}, \Sigma_{1:m|1:m})\right] \\
&= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}_{1:m|1:m}\tilde{\Sigma}_{1:m|1:m}\right) + \frac{1}{2}\ln\left(\frac{\det\Sigma_{1:m|1:m}}{\det\tilde{\Sigma}_{1:m|1:m}}\right) - \frac{k}{2} \\
&\quad + \frac{1}{2}(\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m})^T \Sigma^{-1}_{1:m|1:m}(\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m})
\end{aligned} \tag{18}
$$

We decompose this KL-divergence into two terms.

$$
KL\left[p_{\tilde{K}}(f|y)||p_K(f|y)\right] = \Delta_\Sigma + \Delta_\mu
$$

where

$$
\begin{aligned}
\Delta_\Sigma &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}_{1:m|1:m}\tilde{\Sigma}_{1:m|1:m}\right) + \frac{1}{2}\ln\left(\frac{\det\Sigma_{1:m|1:m}}{\det\tilde{\Sigma}_{1:m|1:m}}\right) - \frac{k}{2} \\
\Delta_\mu &= \frac{1}{2}(\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m})^T \Sigma^{-1}_{1:m|1:m}(\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m})
\end{aligned}
$$

We first consider the Δ_Σ. Unfortunately, it is hard to analytically obtain exact upper bound of this quantity. So we use the Taylor approximation in matrix sense with respect to $\tilde{\Sigma}_{1:m|1:m}$, centered around $\Sigma_{1:m|1:m}$. Using matrix derivative identities $\frac{\partial \log\det X}{\partial X} = X^{-T}$ and $\frac{\partial X^{-1}}{\partial X_{i,j}} = X^{-1}e_i e_j^T X^{-1}$ for square matrix, we get the derivative

$$
\begin{aligned}
\frac{\partial \Delta_\Sigma}{\partial(\tilde{\Sigma}_{1:m|1:m})_{i,j}} &= \frac{1}{2}(\Sigma^{-1}_{1:m|1:m})_{i,j} - \frac{1}{2}(\tilde{\Sigma}^{-1}_{1:m|1:m})_{i,j} \\
\frac{\partial^2 \Delta_\Sigma}{\partial(\tilde{\Sigma}_{1:m|1:m})_{i,j}\partial(\tilde{\Sigma}_{1:m|1:m})_{\tilde{i},\tilde{j}}} &= \frac{1}{2}(\tilde{\Sigma}^{-1}_{1:m|1:m})_{i,j}(\tilde{\Sigma}^{-1}_{1:m|1:m})_{\tilde{i},\tilde{j}}
\end{aligned}
$$

Using the derivatives above, we obtain Taylor approximation

$$
\begin{aligned}
\Delta_\Sigma &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}_{1:m|1:m}(\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m})\Sigma^{-1}_{1:m|1:m}(\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m})\right) + \mathcal{O}(||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}||^3) \\
&\leq \frac{1}{2}\mathrm{tr}\left(\Sigma^{-2}_{1:m|1:m}\right)||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}||^2 + \mathcal{O}(||\tilde{\Sigma}_{1:m|1:m} - \Sigma_{1:m|1:m}||^3) \\
&\leq \frac{9}{2}\mathrm{tr}\left(\Sigma^{-2}_{1:m|1:m}\right)||\tilde{K} - K||^2 + \mathcal{O}(||\tilde{K} - K||^3)
\end{aligned}
$$

We can bound Δ_μ without Taylor expansion as

$$
\begin{aligned}
\Delta_\mu &= \frac{1}{2}(\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m})^T\Sigma^{-1}_{1:m|1:m}(\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m}) \\
&\leq \frac{1}{2}||\mu_{1:m|1:m} - \tilde{\mu}_{1:m|1:m}||_2^2 \cdot ||\Sigma^{-1}_{1:m|1:m}|| \\
&\leq \frac{2||y||_2^2}{\tau^4}||\tilde{K} - K||^2 \cdot ||\Sigma^{-1}_{1:m|1:m}||
\end{aligned}
$$

Thus, combining the upper bounds on the Δ_μ and Δ_Σ, we get

$$
\begin{aligned}
&KL\left[p_{\tilde{K}}(f|y)||p_K(f|y)\right] \\
&\quad = \Delta_\Sigma + \Delta_\mu \\
&\quad \leq \left(\frac{9}{2}\mathrm{tr}\left(\Sigma^{-2}_{1:m|1:m}\right) + \frac{2||y||_2^2}{\tau^4} \cdot ||\Sigma^{-1}_{1:m|1:m}||\right) \cdot ||\tilde{K} - K||^2 + \mathcal{O}(||\tilde{K} - K||^3)
\end{aligned}
$$

.

References

Ailerons dataset. https://sci2s.ugr.es/keel/dataset.php?cod=93. Accessed: 2019-07-05.

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In *Conference on Learning Theory*, pages 103–115, 2016.

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

Mohamed-Ali Belabbas and Patrick J Wolfe. On landmark selection and sampling in high-dimensional data analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4295–4312, 2009a.

Mohamed-Ali Belabbas and Patrick J Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009b.

David R Burt, Carl E Rasmussen, and Mark van der Wilk. Explicit rates of convergence for sparse variational inference in gaussian process regression. *Symposium on Advances in Approximate Bayesian Inference*, 2018.

David R Burt, Carl E Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational gaussian process regression. *arXiv preprint arXiv:1903.03571*, 2019.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556. ACM, 2004.

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

Vincent Fortuin, Gideon Dresdner, Heiko Strathmann, and Gunnar Rätsch. Scalable gaussian processes on discrete domains. *arXiv preprint arXiv:1810.10368*, 2018.

Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast dpp sampling for nyström with application to kernel methods. *arXiv preprint arXiv:1603.06052*, 2016a.

Chengtao Li, Suvrit Sra, and Stefanie Jegelka. Gaussian quadrature for matrix inverse forms with applications. In *International Conference on Machine Learning*, pages 1766–1775, 2016b.

Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Carl E Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. springer, 2006.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009a.

Michalis K Titsias. Variational model selection for sparse gaussian process regression. *Report, University of Manchester, UK*, 2009b.

Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.