Reduced Heteroscedasticity Linear Regression for Nyström Approximation

Hao Yang  
Nanyang Technological University  
Singapore  
lancelot365@gmail.com

Jianxin Wu*  
National Key Laboratory for Novel Software Technology  
Nanjing University, Nanjing 210023, China  
wujx2001@gmail.com

Abstract

The Nyström method is a well known sampling based low-rank matrix approximation approach. It is usually considered to be originated from the numerical treatment of integral equations and eigen-decomposition of matrices. In this paper, we present a novel point of view for the Nyström approximation. We show that theoretically the Nyström method can be regraded as a set of point-wise ordinary least square linear regressions of the kernel matrix, sharing the same design matrix. With the new interpretation, we are able to analyze the approximation quality based on the fulfillment of the homoscedasticity assumption and explain the success and deficiency of various sampling methods. We also empirically show that positively skewed explanatory variable distributions can lead to heteroscedasticity. Based on this discovery, we propose to use non-symmetric explanatory functions to improve the quality of the Nyström approximation with almost no extra computational cost. Experiments show that positively skewed datasets widely exist, and our method exhibits good improvements on these datasets.

1 Introduction

Kernel method is a powerful and important tool for many areas of machine learning, including support vector machines [Cortes and Vapnik, 1995], kernel principle component analysis [Schölkopf et al., 1998], Guassian process [Williams et al., 2002], manifold learning [Talwalkar et al., 2008], etc. These applications usually involve a large kernel matrix that needs up to $O(n^3)$ computations. Large matrices also arises in other applications like clustering and matrix completion. For large scale problems, the entries of these matrices can be in the order of billions, leading to serious difficulties in computing and storing them. Various approximation techniques have been developed in machine learning to solve the problem, e.g., incomplete Cholesky decomposition [Bach and Jordan, 2005], random Fourier features [Rahimi and Recht, 2007] and the Nyström method [Williams and Seeger, 2001].

The Nyström method is a sampling based low-rank matrix approximation method, utilizing the rapidly decaying spectra of kernel matrices. It is usually considered to be originated from the numerical treatment of integral equations and eigen-decomposition of matrices. Sampling techniques play a key role in the Nyström approximation. Therefore, various sampling methods have been developed to improve the approximation quality, to name a few, the diagonal sampling and column-norm sampling that apply a weight to each of sampled column with either diagonal element $K_{ii}$ or the $L_2$ norm of the column [Drineas and Mahoney, 2005; Drineas et al., 2006], adaptive sampling that alternates between selecting a set of columns and updating the distribution over all columns [Deshpande et al., 2006]. Among these methods, K-means sampling [Zhang et al., 2008] shows great results on approximation accuracy but at much higher computational cost [Kumar et al., 2012]. In large scale datasets with high dimensional dense features, the time spent in running the K-means algorithm could be significantly longer than the Nyström approximation itself.

This paper presents a novel point of view for the Nyström approximation, leading to an interpretation of success and deficiency of various sampling methods and a new way to improve the approximation quality. To be specific,

- We demonstrate a novel interpretation of the Nyström method. We show that theoretically the Nyström method can be regraded as a set of point-wise ordinary least square linear regressions of the kernel matrix sharing the same design matrix;
- By employing the linear regression interpretation, we are able to analyze the approximation quality of the Nyström method from a different angle. We show that empirically the heteroscedasticity problem in the ordinary least square linear regression could be used to explain the success and deficiency of various sampling methods, especially the high approximation quality of the K-means sampling method [Zhang et al., 2008];
- We empirically illustrate that positively skewed explanatory variable distributions coincides and may lead to heteroscedasticity. Inspired by this discovery, we can greatly improve the approximation quality of the

\*To whom correspondence should be addressed. This work was done when J. Wu was with the School of Computer Engineering, Nanyang Technological University, Singapore.
Nystro\text{\text{"}}m method on many real world datasets with almost no extra computational cost through a transformation function.

The paper is organized as following: The Nystro\text{\text{"}}m Method and its linear regression interpretation is explained in Section 2. We then demonstrate the heteroscedasticity, its relation to skewed regressor distribution, and propose our solutions in Section 3. The empirical results on real world datasets are presented in Section 4. Finally, we draw conclusions and discuss future directions of the research on Section 5.

2 The Nystro\text{\text{"}}m Method Revisited

The Nystro\text{\text{"}}m method [Williams and Seeger, 2001] provides a low-rank approximation of a kernel matrix based on eigendecomposition and numerical treatment of integral.

Let $K \in \mathbb{R}^{n \times n}$ be a kernel matrix with $K_{i,j} = k(x_i, x_j)$ and $D = \{x_i\}_{i=1}^{n}$. If $C = \{c_i\}_{i=1}^{m}$ is a subset of $D$, which we call anchor points in this paper, the Nystro\text{\text{"}}m method generates an (up to rank $m$) approximation $\bar{K}$ of $K$ by:

$$\bar{K} = SX^+S^T \approx K,$$

(1)

where $S$ is an $n \times m$ matrix with $S_{i,j} = k(x_i, c_j)$; $X$ is an $m \times m$ matrix with $X_{i,j} = k(c_i, c_j)$ and $X^+$ is the Moore-Penrose pseudo-inverse of $X$.

2.1 A Linear Regression Interpretation

The anchor points can be sampled in various fashions, including random sampling [Williams and Seeger, 2001], diagonal sampling [Drineas and Mahoney, 2005], adaptive sampling [Deshpande et al., 2006], $K$-means sampling [Zhang et al., 2008], $K$-means sampling with weight [Zhang and Kwok, 2009], etc. For $K$-means sampling, the anchor points $C$ are chosen as centroids a $K$-means clustering of $D$. These studies [Zhang et al., 2008; Kumar et al., 2012] have shown that $K$-means based sampling provides much better approximation than random sampling at the price of much higher computational cost. In this section, we propose a different interpretation of the Nystro\text{\text{"}}m method, which not only explains the success of $K$-means sampling, but also leads to a computationally efficient improvement of the Nystro\text{\text{"}}m method. Different from the classic eigendecomposition perspective, we interpret the Nystro\text{\text{"}}m method as a set of point-wise linear regressions (sharing the same design matrix) of the kernel matrix.

For a kernel function $k(q,y)$, if we treat $y$ as constant, $k(q,y)$ becomes a function of the variable $q$:

$$f_y(q) = k(q,y).$$

(2)

The generalized multiple linear regression [Gro\text{\text{"}}5, 2003] will estimate the function $f_y(\cdot)$ as:

$$f_y(q) = e(q)^T \beta + \epsilon.$$  

(3)

$f_y(q)$ is the dependent variable or response variable, $e(q)$ is is a vector of the independent variables or explanatory variables, and $\epsilon$ is the error term. We call $e(\cdot)$ the explanatory function. It is a set of basis functions designed to capture the essence of the kernel function $k$. Choosing the appropriate explanatory function $e(\cdot)$ is very important as it determines the quality of approximation.

Given a set of anchor points $C = \{c_i\}_{i=1}^{m}$ for learning the regression models, their true dependent variable values are:

$$s(y) = [k(c_1, y), \ldots, k(c_m, y)]^T.$$  

(4)

It is well known that the ordinary least square (OLS) solution to Equation (3) is:

$$\hat{\beta} = X^+s(y),$$

(5)

where $X$ is called the design matrix and

$$X = (e(c_1), \ldots, e(c_m))^T.$$  

(6)

A natural choice of the explanatory function to capture the characteristic of the kernel function $k$ is to set $e(q) = s(q)$, under which the design matrix becomes

$$X = \begin{bmatrix} k(c_1, c_1) & \ldots & k(c_1, c_m) \\ \vdots & \ddots & \vdots \\ k(c_m, c_1) & \ldots & k(c_m, c_m) \end{bmatrix}.$$  

(7)

Then, $f_y(q)$ is approximated by:

$$f_y(q) = s(q)X^+s(y) + \epsilon.$$  

(8)

Therefore, if we choose the same set of anchor points for all functions $f_{x_1}, \ldots, f_{x_n}$ we want to approximate, for arbitrary $x_i, x_j \in D$:

$$f_{x_i}(x_i) = k(x_i, x_j) = K_{i,j} \approx s(x_i)X^+s(x_j).$$

(9)

Clearly, the kernel matrix $K$ is approximated by:

$$\bar{K} = SX^+S^T,$$  

(10)

in which $S$ is the matrix of similarity functions $s(x_i), i = 1, \ldots, n$, so $S_{i,j} = k(c_i, x_i) = k(x_i, c_j)$. Therefore $\bar{K}$ is exactly the same as the Nystro\text{\text{"}}m approximation result.

The efficiency of the approximation is ensured by employing the same design matrix (i.e., same set of anchor points) for all examples in $D$. For each linear regression, we just change the dependent variable’s true values (i.e., the values of similarity functions). The generality of the approximation is also guaranteed, as in Equation (3), the $y$ value can be chosen arbitrarily from the input feature space.

This new interpretation allows us to have an alternative explanation for the success and deficiency of the Nystro\text{\text{"}}m and improved Nystro\text{\text{"}}m approximation [Zhang et al., 2008]. And, more importantly, we gain the freedom to choose the explanatory function $e(\cdot)$ besides using the same form as the similarity function $s(\cdot)$. We will show in the following sections that the fulfillment for assumptions of OLS, especially homoscedasticity, has obvious impact on the quality of the Nystro\text{\text{"}}m approximation, and by employing non-symmetric explanatory functions, i.e., set $e(\cdot) \neq s(\cdot)$, the approximation quality can be greatly improved without extra computational cost.
3 Heteroscedasticity, Skewness, Their Adverse Effects on Nyström, and Cure

As stated, the Nyström approximation is essentially a set of point-wise OLS linear regressions sharing the same design matrix. The error terms of the anchor points \( \epsilon_i, i = 1, \ldots, m \) are defined as

\[
f_y(c_i) = s(c_i)X^+s(y) + \epsilon_i. \tag{11}
\]

OLS is indeed the best linear unbiased estimator (BLUE) by the Gauss–Markov theorem, given that the errors have expectation zero conditional on the independent variables \((E(\epsilon_i|c) = 0)\), are uncorrelated \((\text{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j)\) and have constant variances \((\epsilon_i \sim N(0|\sigma^2))\) [Groß, 2003]. However, if these assumptions are violated, the robustness of OLS can be doubtful.

The situation where the constant variance assumption (i.e., homoscedastic) is violated is called heteroscedastic. In many real world datasets, we observe that heteroscedasticity indeed exists and hurts the quality of the Nyström approximation if random sampling scheme is applied. Therefore, we mainly deal with the heteroscedasticity problem in this paper.

3.1 Heteroscedasticity and Its Detection

Heteroscedasticity refers to the situation where the error terms have non-constant variances. The presence of heteroscedasticity invalidates many statistical tests of significance. Thus it is a major concern in the application of regression analysis. In the linear regression model, a test for heteroscedasticity is a test of the null hypothesis [Groß, 2003]

\[ H_0 : \text{Var}(\epsilon_i) = \sigma^2, \text{ for } i = 1, \ldots, m. \tag{12} \]

Here the error \( \epsilon_i \) is the \( i \)-th element of the vector of least squares residuals \( f_y - CX \) (also cf. Equation (11)). The alternative hypothesis is rather unspecified. Several statistical tests can be applied to detect the existence of heteroscedasticity, including the White test and the Breusch-Pagan test. In this paper, we choose the Breusch-Pagan test as the tool, in which the alternative \( H_1 \) is expressed as [Breusch and Pagan, 1979]:

\[ H_1 : \text{Var}(\epsilon_i) = h(\alpha_0 + z_i^T \alpha), \tag{13} \]

where \( z_i^T, i = 1, \ldots, m \), is the explanatory variables and \( \alpha \in \mathbb{R}^d \) is unknown parameters vector, \( q \) is the dimension of the explanatory variable. The function \( h(\cdot) \) needs not to be specified. The Breusch-Pagan test performs a series of auxiliary regressions to get the test statistics \( Q \), which should be asymptotically \( \chi^2(p) \) distributed under \( H_0 \). The hypothesis \( H_0 \) is rejected at level \( \alpha \) if \( Q > \chi^2(p,1-\alpha) \). We use the \( p \)-value of the \( \chi^2 \) distribution as the indicated value of homoscedasticity; the higher \( p \)-value we get from the Breusch-Pagan test, the higher probability \( Q \) is asymptotically \( \chi^2(p) \) distributed under \( H_0 \), the more homoscedastic the regression model is. We will show that, starting with synthetic datasets, severe heteroscedasticity problem exists on many kinds of data distributions if the anchor points \( C \) are chosen randomly, and this problem will hurt the linear regression, and consequently the Nyström method.

We introduce two synthetic datasets to show the existence of heteroscedasticity in the Nyström method with randomly sampled anchor points. We generated 1000 training instances, each instance consists of 100 observations sampled from a univariate Gaussian distribution or a lognormal distribution. The Radial Basis Function \( k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \gamma) \) is used as the kernel function, with \( \gamma \) set to be the average square distance between training instances, following [Zhang et al., 2008].

The Breusch-Pagan test is used to test the heteroscedasticity of the Nyström approximation on these two synthetic datasets. The homoscedasticity is measured by \( p \)-value, and the approximation error (shown in italic inside the parenthesis) is measured by

\[
\text{error} = \left\| \mathbf{K} - \tilde{\mathbf{K}} \right\|_F, \tag{14}
\]

where \( \left\| \cdot \right\|_F \) is the Frobenius norm. For both random sampling and \( K \)-means sampling, we sampled 100 training instances as anchor points. We repeat the approximation experiment and the test 10 times and use the average values as the final results. Due to space limit, we omit the standard deviations of approximation errors since they are small.

As shown in Table 1, the Nyström method with random sampling has poor homoscedasticity (below 0.5) while the \( K \)-means Nyström has very good homoscedasticity (about 0.9), leading to better approximation results. We also observe the same phenomenon in most real world datasets, which we will present in details in Section 4.

3.2 Heteroscedasticity, Skewness, and Cure

A variety of reasons could account for the heteroscedasticity, including aggregated and grouped data, the presence of outliers [Groß, 2003], the regression model is not correctly specified, and the skewness in the distribution of one or more regressors (explanatory variables) in the model [Fox, 1997].

One common way to overcome heteroscedasticity is to use weighted ordinary least square (WOLS) estimation for linear regression models. The WOLS model is [Groß, 2003]:

\[
\hat{\beta}_W = (X^TWX)^{-1}X^TWs(y). \tag{15}
\]

Here \( W \) is a diagonal weight matrix estimated by the error term \( \epsilon_i \). Unfortunately, if we apply WOLS to solve the heteroscedasticity problem in the Nyström case, we have to estimate a different \( W \) for every \( x_i \in D \) (cf. Equation (5)), which will lead to \( O(mn^2) \) more computations. Therefore, it is not feasible to apply WOLS here unless a universally valid \( W \) for the whole dataset exists and can be efficiently found.

Alternatively, we conjecture that in the Nyström method, positively skewed explanatory variables will cause severe heteroscedasticity. Skewness is a measure of the asymmetry of the probability distribution of a random variable. It is denoted as \( \gamma_1 \) and defined as \( \gamma_1 = E \left[ \frac{(S - \mu)^3}{\sigma^3} \right] = \frac{\mu_3}{\sigma^3} \)

\[
= \frac{E \left[ (S - \mu)^3 \right]}{\left( E \left[ (S - \mu)^2 \right] \right)^{3/2}} = \frac{\mu_3}{\sigma^3}, \text{ where } \mu_3 \text{ is the third moment about the mean } \mu, \sigma \text{ is the standard deviation, } S \text{ is the explanatory variables (the first term of the right-hand side of Equation (10)), and } E \text{ is the expectation operator.}
named the power transformation. We utilize this well-known
1977], especially a family of rank-preserving transformation
propose to use non-symmetric explanatory function to solve
the anchor points affects the distributions of explanatory vari-
ares more evenly distributed, thus the explanatory variable dis-
K-points are chosen as the centroids of
which may cause positively skewed explanatory variable dis-
umpredictable and could be in the outskirts of the datasets,
 approximation.
For random sampling, the position of the anchor points is
unpredictable and could be in the outskirts of the datasets,
which may cause positively skewed explanatory variable dis-
tributions on many datasets since kernels are mostly distance
based (e.g. RBF kernel). On the other hand, if the anchor
points are chosen as the centroids of K-means clusters, they
are more evenly distributed, thus the explanatory variable dis-
tributions would be less biased. The fact that positions of
the anchor points affects the distributions of explanatory vari-
ables and positively skewed distributions cause homoscedas-
ticy explain why K-means sampling has better homoscedas-
ticy.

With the linear regression interpretation, we are able to
propose to use non-symmetric explanatory function to solve
the heteroscedasticity problem without changing the sam-
ping scheme. In the statistical literature, data transforma-
tion is widely used to cure the positively skewed data [Tukey,
1977], especially a family of rank-preserving transformation
named the power transformation. We utilize this well-known
fact, and design an innovative explanatory function:
\[
\hat{e}(x) = [1, T(k(c_1, x)), \ldots, T(k(c_m, x))]^T. \quad (16)
\]
Here T(·) is the transformation function and 1 is a bias term
for further stabilizing the variance of explanatory variables.
Correspondingly, the design matrix is then:
\[
\tilde{X} = \begin{bmatrix}
1 & T(k(c_1, c_1)) & \cdots & T(k(c_1, c_m)) \\
\vdots & \vdots & \ddots & \vdots \\
1 & T(k(c_m, c_1)) & \cdots & T(k(c_m, c_m)) 
\end{bmatrix}. \quad (17)
\]
Then the kernel matrix is approximated by:
\[
\tilde{K}_T = \tilde{E}\tilde{X}^T S^T \approx K, \quad (18)
\]
where \(\tilde{E}\) is the matrix of the new explanatory functions
\(\hat{e}(x)_i, i = 1, \ldots, n\). Notice that all the linear regression func-
tions still share the same design matrix, thus the efficiency
of the Nyström method is preserved. Since our explanatory
function is not the same as the similarity function, the
approximation matrix \(\tilde{K}_T\) is non-symmetric. Therefore, we use

As shown in Table 2, positively skewed explanatory vari-
ables distributions tend to coexist with heteroscedasticity.
Empirical results of many real world datasets also prove that,
illustrated in Figure 1. When the skewness of explanatory
variables is higher than certain threshold, the homoscedas-
ticity drops remarkably. Empirically, we find out that explana-
tory variable distributions with skewness higher than 1.5 will
have the heteroscedastic problem, and homoscedasticity as-
sumption is satisfied if the p-value is above 0.7 in the Nyström
approximation.

\begin{table}[h]
\centering
\caption{Homoscedasticity (Approximation Error) of two Synthetic Datasets}
\begin{tabular}{|c|c|c|c|}
\hline
 & Random & K-means & $T_1$ & $T_2$ \\
\hline
Normal & 0.43 ± 0.19 (31.34) & 0.92 ± 0.01 (26.33) & 0.57 ± 0.16 (29.66) & 0.76 ± 0.06 (26.33) \\
Lognormal & 0.45 ± 0.16 (31.08) & 0.90 ± 0.03 (26.68) & 0.54 ± 0.14 (29.42) & 0.73 ± 0.05 (27.78) \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Homoscedasticity (Skewness of Explanatory Variables) of two Synthetic Datasets}
\begin{tabular}{|c|c|c|c|}
\hline
 & Random & K-means & $T_1$ & $T_2$ \\
\hline
Normal & 0.43 ± 0.19 (5.81) & 0.92 ± 0.01 (3.28) & 0.57 ± 0.16 (3.22) & 0.76 ± 0.06 (0.21) \\
Lognormal & 0.45 ± 0.16 (3.23) & 0.90 ± 0.03 (0.23) & 0.54 ± 0.14 (1.51) & 0.73 ± 0.05 (0.50) \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{relationship.png}
\caption{The Relationship between Homoscedasticity and Skewness}
\end{figure}

a simple method to symmetries it:
\[
\tilde{K}_T = \frac{K_T + K_T^T}{2}, \quad (19)
\]
We tested two transformation functions: $T_1(x) = \ln(1+x)$
and $T_2(x) = x^{1/2}$. They both are effective on curing the pos-
itively skewed data and make it more normal distribution-like.
As a result, the heteroscedasticity problem is greatly relieved
and the approximation quality is improved. The effects of our
method can be demonstrated in the two synthetic datasets as
show in Tables 1 and 2. Clearly, both $T_1$ and $T_2$ have good
effects on correcting the positively skewed data, and since $T_2$
are more powerful in the correction, it has better approxima-
tion quality (almost as good as K-means) than $T_2$. Details of
our method can be found in Algorithm 1. We will show the
experimental results on real world datasets in Section 4.

\section{Experimental Results}
In this section we presents empirical evaluations of our trans-
formation methods. We mainly examine the performance of
the low-rank approximation methods by measuring their ap-
proximation errors. The methods we tested are the Nyström
method, K-means sampling Nyström, and $T_1$ and $T_2$ trans-
formed Nyström. The computational complexities of differ-
ent method are summarized in Table 3, where $d$ is average
non-zero dimensions of the dataset. Since high dimensional
data naturally arises from many applications, such as com-
puter vision, $d$ is likely to be much larger than $m$, leading
to extremely expensive K-means cost. For our method, the
only extra computation cost is the transformation function
$T$, which is just $O(nm)$, much cheaper than K-means sam-
ping and WOLS. We also shown the real CPU time for the
Algorithm 1 Reduced Heteroscedasticity Linear Regression Nyström Approximation

1: Given a dataset \( D = \{x_i\}_{i=1}^n \), randomly sample \( m \) anchor points as \( D = \{c_i\}_{i=1}^m \).
2: Calculate explanatory matrix \( E \) (which is the same as \( S \)) and \( X \).
3: Calculate the skewness of \( E \), denoted as \( SK_E \).
4: if \( SK_E > \text{thres} \) then
5: Apply transformation function \( T_1 \) or \( T_2 \) to \( E \) and \( X \) to get \( \tilde{E} \) and \( \tilde{X} \).
6: The approximation is then \( \tilde{K}_T = \tilde{E} \tilde{X}^T S^T \).
7: Symmetrize \( \tilde{K}_T \) with Equation (19).
8: else
9: Use the original Nyström approximation \( \tilde{K} = E X^T S^T \).
10: end if
11: Output: The approximated matrix \( \tilde{K}_T \) or \( \tilde{K} \).

scene15 dataset, a moderate size vision dataset, with 100 anchor points. Our method has almost the same computational time as random sampling, while \( K \)-means is about 10 times slower.

We choose a number of benchmark datasets from the LIBSVM archive. The vision dataset scene15 is from [Lazebnik et al., 2006]. We generate the dataset using Bag-of-Words model and Spatial Pyramid with libHIK [Wu, 2010]. The datasets summary is shown in Table 5. Gaussian kernel 
\[ k(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{\gamma}) \]

is used here with \( \gamma \) set to be the average square distance between training instances. To produce statistically reliable results, we repeat all the methods 10 times to get the mean and standard deviations.

The approximation errors are shown in Figure 2. The \( x \)-axis shows the ratio of \( m/n \), i.e., the sample rate, which we choose from 0.01 to 0.1, following [Zhang et al., 2008]. The \( y \)-axis shows the approximation errors as defined in Equation (14). The experiments demonstrate that \( K \)-means sampling still provides the best approximation results. However, as we have discussed before, \( K \)-means clustering itself is very expensive and can be much slower than the Nyström method. For large scale problem, it is not even feasible to perform \( K \)-means clustering [Kumar et al., 2012]. Our methods are shown as LogNyström and SqrtNyström, corresponding to the \( T_1 \) and \( T_2 \) transformation functions. These two transformations both show good improvements on the random sampling with almost no extra cost, especially the SqrtNyström. For several datasets, the SqrtNyström exhibits comparable results to \( K \)-means sampling. We have also tested a simple implementation of WOLS. The weight matrix is evaluated from error terms of each linear regression. Preliminary

Table 3: Computational Cost Compared to Random Sampling

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Transformations</th>
<th>( K )-means</th>
<th>WOLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional Complexities</td>
<td>0</td>
<td>( O(mn) )</td>
<td>( O(mnd) )</td>
<td>( O(mn^2) )</td>
</tr>
<tr>
<td>CPU Time (scene15)</td>
<td>1.00×</td>
<td>1.10×</td>
<td>9.55×</td>
<td>51.50×</td>
</tr>
</tbody>
</table>

Table 5: Summary of datasets

<table>
<thead>
<tr>
<th></th>
<th>germain</th>
<th>splice</th>
<th>adult1a</th>
<th>dna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>1000</td>
<td>1000</td>
<td>1605</td>
<td>2000</td>
</tr>
<tr>
<td>Dimension</td>
<td>24</td>
<td>60</td>
<td>123</td>
<td>180</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>mushrooms</th>
<th>sonar</th>
<th>mnist</th>
<th>scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>8124</td>
<td>208</td>
<td>1605</td>
<td>1500</td>
</tr>
<tr>
<td>Dimension</td>
<td>112</td>
<td>60</td>
<td>123</td>
<td>12400</td>
</tr>
</tbody>
</table>
results show that WOLS is more costly than K-means and inferior in accuracy, making it unsuitable to be used in practical situations.

We again examine the relationship between homoscedasticity, skewness and the approximation quality. We sampled 100 anchor points from each dataset to do the Breusch-Pagan test and repeat 10 times for each method. The results shown in Table 4 again supports our analysis: Positively skewed explanatory variable distributions from random sampling scheme leads to heteroscedasticity, and thus affects the approximation quality. Apart from sampling good anchor points using K-means clustering to avoid the skewness problem, applying transformation functions can serve well for the purpose, too, with almost no cost.

4.1 Failure Mode Analysis

Our transformation works well for positively skewed explanatory variable distributions, which widely exists in real world datasets if random sampling scheme is applied. However, for datasets with normal-like explanatory variable distributions or if anchor points are K-means sampled, our method may not have positive effects, because they do not suffer from the heteroscedasticity problem.

We have tested over 16 datasets, among which 6 are without the heteroscedasticity problem. An example of such datasets is svmguide1, whose approximation results are illustrated in Figure 3. We do not show others due to space limit. As shown in Table 4, the distribution of explanatory variables is not positively skewed and there is no heteroscedasticity issue. Therefore, the transformations do not improve the approximation accuracy. K-means sampling, on the other hand, still have positive effect, but the improvement is not as significant as on the heteroscedastic datasets. This is also an evidence for our reasoning for the success of K-means sampling.

Discussions

Learning from the failure cases, we suggest to calculate the skewness of explanatory variable distribution before applying data transformations. Since calculating skewness only involve the kernel values, which are just scalars, it is also very cheap compared to K-means clustering. As we have discussed in previous section, the empirical threshold we suggest is 1.5. By setting up the threshold, we can successfully avoid the failure cases and make good use of the transformation to get better approximation results, thus Algorithm 1 can be safely applied on almost all datasets.

5 Conclusions

In this paper, we present a linear regression interpretation for the Nyström approximation and propose to use non-symmetric explanatory functions to improve the quality of the Nyström approximation with almost no extra computational cost. To be specific, we show that theoretically the Nyström method can be regarded as a set of point-wise ordinary least square linear regressions of the kernel matrix, sharing the same design matrix. We analyze the approximation quality based on the fulfillment of the homoscedasticity assumption and explain the success and deficiency of random and K-means sampling. Empirically, we demonstrate that positively skewed explanatory variable distributions can lead to heteroscedasticity, thus we propose to use the transformation functions to normalize the skewed distribution and overcome heteroscedasticity. In large scale problems, when K-means clustering is not feasible, our method can still be applied efficiently. The experimental results on various real world datasets prove the effectiveness of our method.

In the future, we want to explore the possibility to apply transformations to the ensemble Nyström method [Kumar et al., 2012] to further improve our method. The ensemble Nyström method combines multiple Nyström approximations, which is called expert, with several weight functions. If we apply data transformation to each expert and also combine them with appropriate weight function, we expect to get better results on heteroscedastic datasets. We also want to found a theoretical analysis on the relationship between positively skewed regressors and heteroscedasticity so that our method is more theoretically sound.
References


