# 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 eigendecomposition 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 eigendecomposition 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 $\mathbf{K}_{ii}$ 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.

Nyström method on many real world datasets with almost no extra computational cost through a transformation function.

The paper is organized as following: The Nyström Method and its linear regression interpretation is explained in Section 2. We then demonstrate the heterscedasticity, 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 Nyström Method Revisited

The Nyström method [Williams and Seeger, 2001] provides a low-rank approximation of a kernel matrix based on eigendecomposition and numerical treatment of integral.

Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be a kernel matrix with $\mathbf{K}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\mathbf{D} = \{\boldsymbol{x}_i\}_{i=1}^n$. If $\mathbf{C} = \{\boldsymbol{c}_i\}_{i=1}^m$ is a subset of $\mathbf{D}$, which we call *anchor points* in this paper, the Nyström method generates an (up to rank $m$) approximation $\widetilde{\mathbf{K}}$ of $\mathbf{K}$ by:

$$\widetilde{\mathbf{K}} = \mathbf{S}\mathbf{X}^+\mathbf{S}^T \approx \mathbf{K}, \tag{1}$$

where $\mathbf{S}$ is an $n \times m$ matrix with $\mathbf{S}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{c}_j)$; $\mathbf{X}$ is an $m \times m$ matrix with $\mathbf{X}_{i,j} = k(\boldsymbol{c}_i, \boldsymbol{c}_j)$ and $\mathbf{X}^+$ is the Moore-Penrose pseudo-inverse of $\mathbf{X}$.

### 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 $\mathbf{C}$ are chosen as centroids a $K$-means clustering of $\mathbf{D}$. 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 Nyström method, which not only explains the success of $K$-means sampling, but also leads to a computationally efficient improvement of the Nyström method. Different from the classic eigendecomposition perspective, we interpret the Nyström method as *a set of point-wise linear regressions (sharing the same design matrix) of the kernel matrix.*

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

$$f_{\boldsymbol{y}}(\boldsymbol{q}) = k(\boldsymbol{q}, \boldsymbol{y}). \tag{2}$$

The generalized multiple linear regression [Groß, 2003] will estimate the function $f_{\boldsymbol{y}}(\cdot)$ as:

$$f_{\boldsymbol{y}}(\boldsymbol{q}) = e(\boldsymbol{q})^T\boldsymbol{\beta} + \epsilon. \tag{3}$$

$f_{\boldsymbol{y}}(\boldsymbol{q})$ is the dependent variable or response variable, $e(\boldsymbol{q})$ 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 $\mathbf{C} = \{\boldsymbol{c}_i\}_{i=1}^m$ for learning the regression models, their true dependent variable values are:

$$s(\boldsymbol{y}) = [k(\boldsymbol{c}_1, \boldsymbol{y}), \ldots, k(\boldsymbol{c}_m, \boldsymbol{y})]^T. \tag{4}$$

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

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^+ s(\boldsymbol{y}), \tag{5}$$

where $\mathbf{X}$ is called the design matrix and

$$\mathbf{X} = (e(\boldsymbol{c}_1), \ldots, e(\boldsymbol{c}_m))^T. \tag{6}$$

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

$$\mathbf{X} = \begin{bmatrix} k(\boldsymbol{c}_1, \boldsymbol{c}_m) & \ldots & k(\boldsymbol{c}_1, \boldsymbol{c}_m) \\ \vdots & \vdots & \vdots \\ k(\boldsymbol{c}_m, \boldsymbol{c}_1) & \ldots & k(\boldsymbol{c}_m, \boldsymbol{c}_m) \end{bmatrix}. \tag{7}$$

Then, $f_{\boldsymbol{y}}(\boldsymbol{q})$ is approximated by:

$$f_{\boldsymbol{y}}(\boldsymbol{q}) = s(\boldsymbol{q})\mathbf{X}^+ s(\boldsymbol{y}) + \epsilon. \tag{8}$$

Therefore, if we choose the same set of anchor points for all functions $f_{\boldsymbol{x}_1}, \ldots, f_{\boldsymbol{x}_n}$ we want to approximate, for arbitrary $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbf{D}$:

$$f_{\boldsymbol{x}_j}(\boldsymbol{x}_i) = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbf{K}_{i,j} \approx s(\boldsymbol{x}_i)\mathbf{X}^+ s(\boldsymbol{x}_j). \tag{9}$$

Clearly, the kernel matrix $\mathbf{K}$ is approximated by:

$$\widetilde{\mathbf{K}} = \mathbf{S}\mathbf{X}^+\mathbf{S}^T, \tag{10}$$

in which $\mathbf{S}$ is the matrix of similarity functions $s(x_i), i = 1, \ldots, n$, so $\mathbf{S}_{i,j} = k(\boldsymbol{c}_j, \boldsymbol{x}_i) = k(\boldsymbol{x}_i, \boldsymbol{c}_j)$. Therefore $\widetilde{\mathbf{K}}$ is exactly the same as the Nyström 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 $\mathbf{D}$. 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 $\boldsymbol{y}$ 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 Nyström and improved Nyström 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 Nyström 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 Heterscedasticity, 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

$$f_{\boldsymbol{y}}(\boldsymbol{c}_i) = s(\boldsymbol{c}_i)\mathbf{X}^{+}s(\boldsymbol{y}) + \epsilon_i. \tag{11}$$

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 ($E(\epsilon_i|\boldsymbol{c}) = 0$), 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 $f_{\boldsymbol{y}} - \mathbf{C}\hat{\boldsymbol{\beta}}$ (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]:

$$H_1 : \text{Var}(\epsilon_i) = h(\alpha_0 + \boldsymbol{z}_i^T\boldsymbol{\alpha}), \tag{13}$$

where $\boldsymbol{z}_i^T, i = 1, \ldots, m$, is the explanatory variables and $\boldsymbol{\alpha} \in \mathbb{R}^q$ 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 $\alpha$ is $Q > \chi^2_{(p),1-\alpha}$. 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 $\mathbf{C}$ 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 $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2/\gamma)$ 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 parentheses) is measured by

$$\text{error} = \left\|\mathbf{K} - \widetilde{\mathbf{K}}\right\|_F, \tag{14}$$

where $\|\cdot\|_F$ 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]:

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}s(\boldsymbol{y}). \tag{15}$$

Here $\mathbf{W}$ 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 $\mathbf{W}$ for every $\boldsymbol{x}_i \in \mathbf{D}$ (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 $\mathbf{W}$ 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 real-valued random variable. It is denoted as $\gamma_1$ and defined as $\gamma_1 = \text{E}\left[\left(\frac{\mathbf{S}-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\text{E}\left[(\mathbf{S}-\mu)^3\right]}{(\text{E}\left[(\mathbf{S}-\mu)^2\right])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$ where $\mu_3$ is the third moment about the mean $\mu$, $\sigma$ is the standard deviation, $\mathbf{S}$ is the explanatory variables (the first term of the right-hand side of Equation (10)), and E is the expectation operator.

Table 1: Homoscedasticity (*Approximation Error*) of two Synthetic Datasets

| | Random | $K$-means | $\mathrm{T}_1$ | $\mathrm{T}_2$ |
|---|---|---|---|---|
| Normal | $0.43 \pm 0.19(31.34)$ | $0.92 \pm 0.01(26.33)$ | $0.57 \pm 0.16(29.66)$ | $0.76 \pm 0.06(26.35)$ |
| Lognormal | $0.45 \pm 0.16(31.08)$ | $0.90 \pm 0.03(26.68)$ | $0.54 \pm 0.14(29.42)$ | $0.73 \pm 0.05(27.78)$ |

Table 2: Homoscedasticity (*Skewness of Explanatory Variables*) of two Synthetic Datasets

| | Random | $K$-means | $\mathrm{T}_1$ | $\mathrm{T}_2$ |
|---|---|---|---|---|
| Normal | $0.43 \pm 0.19(5.81)$ | $0.92 \pm 0.01(0.28)$ | $0.57 \pm 0.16(3.22)$ | $0.76 \pm 0.06(0.21)$ |
| Lognormal | $0.45 \pm 0.16(3.23)$ | $0.90 \pm 0.03(0.23)$ | $0.54 \pm 0.14(1.51)$ | $0.73 \pm 0.05(0.50)$ |

As shown in Table 2, positively skewed explanatory variables 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 homoscedasticity drops remarkably. Empirically, we find out that explanatory variable distributions with skewness higher than 1.5 will have the heteroscedastic problem, and homoscedasticy assumption is satisfied if the $p$-value is above 0.7 in the Nyström 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 distributions 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 distributions would be less biased. The fact that positions of the anchor points affects the distributions of explanatory variables and positively skewed distributions cause heteroscedasticity explain why $K$-means sampling has better homoscedasticity.

With the linear regression interpretation, we are able to propose to use non-symmetric explanatory function to solve the heteroscedasticity problem without changing the sampling scheme. In the statistical literature, data transformation 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:

$$\hat{e}(\boldsymbol{x}) = [1, \mathrm{T}(k(\boldsymbol{c}_1, \boldsymbol{x})), \ldots, \mathrm{T}(k(\boldsymbol{c}_m, \boldsymbol{x}))]^T. \tag{16}$$

Here $\mathrm{T}(\cdot)$ is the transformation function and 1 is a bias term for further stabilizing the variance of explanatory variables. Correspondingly, the design matrix is then:

$$\widehat{\mathbf{X}} = \begin{bmatrix} 1 & \mathrm{T}(k(\boldsymbol{c}_1, \boldsymbol{c}_1)) & \ldots & \mathrm{T}(k(\boldsymbol{c}_1, \boldsymbol{c}_m)) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{T}(k(\boldsymbol{c}_m, \boldsymbol{c}_1)) & \ldots & \mathrm{T}(k(\boldsymbol{c}_m, \boldsymbol{c}_m)) \end{bmatrix}. \tag{17}$$

Then the kernel matrix is approximated by:

$$\widetilde{\mathbf{K}}_{\mathrm{T}} = \widehat{\mathbf{E}}\widehat{\mathbf{X}}^{+}\mathbf{S}^T \approx \mathbf{K}, \tag{18}$$

where $\widehat{\mathbf{E}}$ is the matrix of the new explanatory functions $\hat{e}(\boldsymbol{x}_i), i = 1, \ldots, n$. Notice that all the linear regression functions 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 $\widetilde{\mathbf{K}}_{\mathrm{T}}$ is non-symmetric. Therefore, we use

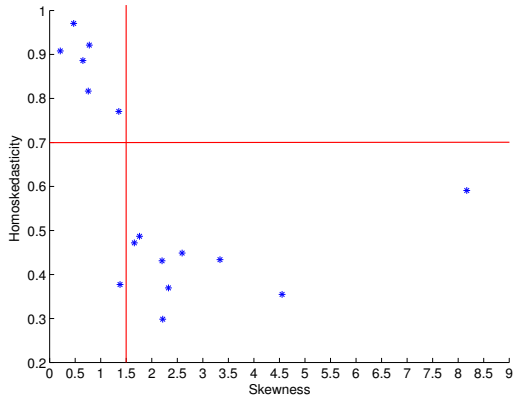



Figure 1: The Relationship between Homoscedasticity and Skewness

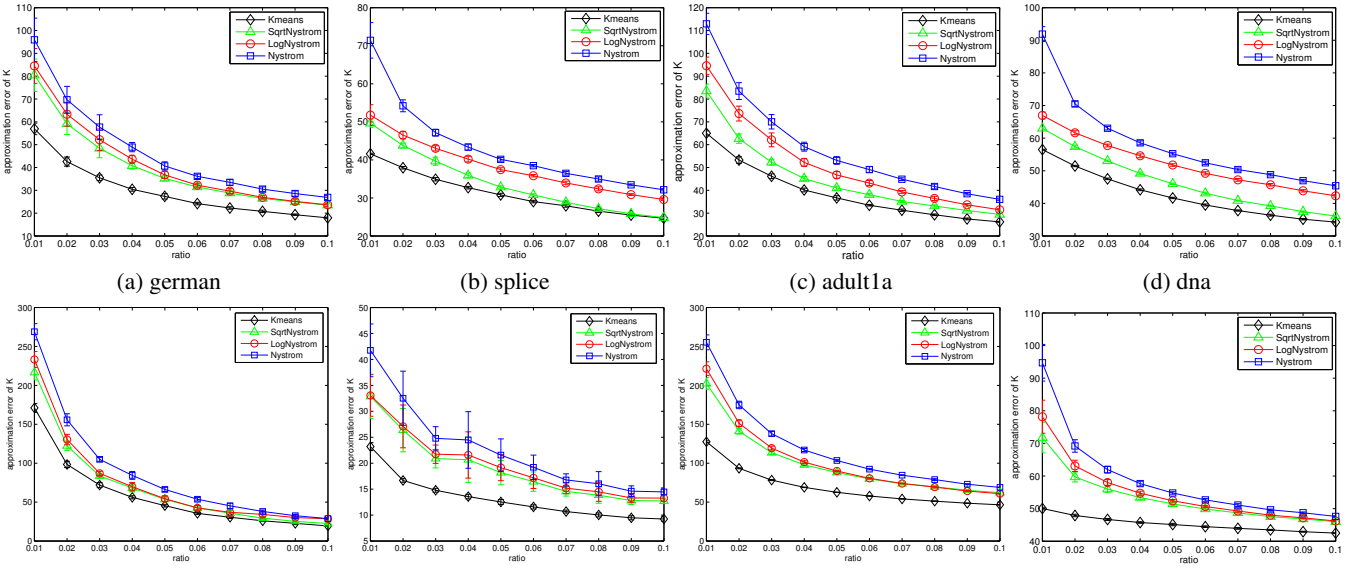
a simple method to symmetries it:

$$\widetilde{\mathbf{K}}_{\mathrm{T}} = \frac{\widetilde{\mathbf{K}}_{\mathrm{T}} + \widetilde{\mathbf{K}}_{\mathrm{T}}^T}{2}, \tag{19}$$

We tested two transformation functions: $\mathrm{T}_1(x) = \ln(1+x)$ and $\mathrm{T}_2(x) = x^{1/2}$. They both are effective on curing the positively 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 $\mathrm{T}_1$ and $\mathrm{T}_2$ have good effects on correcting the positively skewed data, and since $\mathrm{T}_2$ are more powerful in the correction, it has better approximation quality (almost as good as $k$-means) than $\mathrm{T}_2$. Details of our method can be found in Algorithm 1. We will show the experimental results on real world datasets in Section 4.

## 4 Experimental Results

In this section we presents empirical evaluations of our transformation methods. We mainly examine the performance of the low-rank approximation methods by measuring their approximation errors. The methods we tested are the Nyström method, $K$-means sampling Nyström, and $\mathrm{T}_1$ and $\mathrm{T}_2$ transformed Nyström. The computational complexities of different 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 computer 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 sampling and WOLS. We also shown the real CPU time for the




(a) german (b) splice (c) adult1a (d) dna

(e) mushrooms (f) sonar (g) mnist (h) scene

Figure 2: Approximation errors on the kernel matrix

Table 3: Computational Cost Compared to Random Sampling

| | Random | Transformations | $K$-means | WOLS |
|---|---|---|---|---|
| Additional Complexities | 0 | $O(mn)$ | $O(mnd)$ | $O(mn^2)$ |
| CPU Time (*scene15*) | 1.00× | 1.10× | 9.55× | 51.50× |

**Algorithm 1** Reduced Heteroscedasticity Linear Regression Nyström Approximation

1: Given a dataset $\mathbf{D} = \{\boldsymbol{x}_i\}_{i=1}^n$, randomly sample $m$ anchor points as $\mathbf{D} = \{\boldsymbol{c}_i\}_{i=1}^m$
2: Calculate explanatory matrix $\mathbf{E}$ (which is the same as $\mathbf{S}$) and $\mathbf{X}$.
3: Calculate the skewness of $\mathbf{E}$, denoted as $SK_E$.
4: **if** $SK_E > thres$ **then**
5: Apply transformation function $\mathrm{T}_1$ or $\mathrm{T}_2$ to $\mathbf{E}$ and $\mathbf{X}$ to get $\widehat{\mathbf{E}}$ and $\widehat{\mathbf{X}}$.
6: The approximation is then $\widetilde{\mathbf{K}}_{\mathrm{T}} = \widehat{\mathbf{E}}\widehat{\mathbf{X}}^{+}\mathbf{S}^T$.
7: Symmetrize $\widetilde{\mathbf{K}}_{\mathrm{T}}$ with Equation (19).
8: **else**
9: Use the original Nyström approximation $\widetilde{\mathbf{K}} = \mathbf{E}\mathbf{X}^{+}\mathbf{S}^T$.
10: **end if**
11: **Output**: The approximated matrix $\widetilde{\mathbf{K}}_{\mathrm{T}}$ or $\widetilde{\mathbf{K}}$ .

*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.[1] 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 $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / \gamma)$ 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.

Table 5: Summary of datasets

| Datasets | german | splice | adult1a | dna |
|---|---|---|---|---|
| Size | 1000 | 1000 | 1605 | 2000 |
| Dimension | 24 | 60 | 123 | 180 |
| Datasets | mushrooms | sonar | mnist | scene |
| Size | 8124 | 208 | 1605 | 1500 |
| Dimension | 112 | 60 | 123 | 12400 |

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 $\mathrm{T}_1$ and $\mathrm{T}_2$ 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 results show that WOLS is more costly than $K$-means and in-

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Table 4: Homoscedasticity (*Skewness*) of Datasets

| | Random | LogNyström | SqrtNyström | $K$-means |
|---|---|---|---|---|
| german | $0.49 \pm 0.17(1.76)$ | $0.69 \pm 0.16(1.34)$ | $0.93 \pm 0.02(0.03)$ | $0.75 \pm 0.05(1.39)$ |
| splice | $0.43 \pm 0.22(3.34)$ | $0.62 \pm 0.24(1.90)$ | $0.82 \pm 0.07(-0.01)$ | $0.90 \pm 0.03(0.22)$ |
| adult1a | $0.37 \pm 0.11(2.33)$ | $0.63 \pm 0.10(1.76)$ | $0.79 \pm 0.05(1.13)$ | $0.89 \pm 0.05(1.42)$ |
| dna | $0.59 \pm 0.30(8.17)$ | $0.61 \pm 0.31(4.65)$ | $0.83 \pm 0.17(0.80)$ | $0.97 \pm 0.02(0.038)$ |
| mushrooms | $0.43 \pm 0.22(2.20)$ | $0.50 \pm 0.25(1.93)$ | $0.76 \pm 0.08(1.56)$ | $0.92 \pm 0.02(1.99)$ |
| sonar | $0.47 \pm 0.12(1.66)$ | $0.72 \pm 0.11(1.18)$ | $0.87 \pm 0.05(0.50)$ | $0.56 \pm 0.05(0.82)$ |
| mnist | $0.14 \pm 0.22(2.08)$ | $0.31 \pm 0.11(1.55)$ | $0.78 \pm 0.11(0.84)$ | $0.87 \pm 0.02(1.40)$ |
| scene | $0.38 \pm 0.30(1.38)$ | $0.48 \pm 0.17(0.69)$ | $0.65 \pm 0.08(-0.07)$ | $0.92 \pm 0.03(0.59)$ |
| svmguide1 | $0.91 \pm 0.03(0.21)$ | $0.55 \pm 0.14(0.01)$ | $0.57 \pm 0.13(-0.39)$ | $0.91 \pm 0.02(0.25)$ |

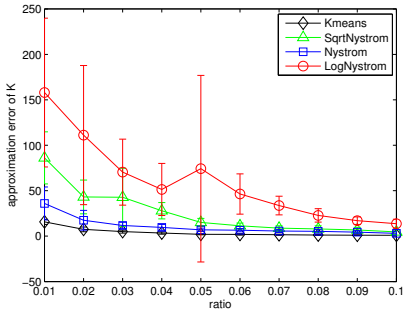



Figure 3: Approximation errors on svmguide1

ferior in accuracy, making it unsuitable to be used in practical situations.

We again examine the relationship between homoscedasticy, 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 heteroscedasticy 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 heteroscedasticy 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 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. 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


[Bach and Jordan, 2005] Francis R. Bach and Michael I. Jordan. Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd International Conference on Machine learning*, pages 33–40, 2005.

[Breusch and Pagan, 1979] Trevor S. Breusch and Adrian Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294, 1979.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[Deshpande *et al.*, 2006] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1117–1126, New York, NY, USA, 2006.

[Drineas and Mahoney, 2005] Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[Drineas *et al.*, 2006] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal of Computing*, 36(1):158–183, July 2006.

[Fox, 1997] John Fox. *Applied Regression Analysis, Linear Models, and Related Methods*. SAGE Publications, 1997.

[Groß, 2003] Jürgen Groß. *Linear Regression*. Springer, 2003.

[Kumar *et al.*, 2012] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 14:981–1006, June 2012.

[Lazebnik *et al.*, 2006] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[Rahimi and Recht, 2007] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184, 2007.

[Schölkopf *et al.*, 1998] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

[Talwalkar *et al.*, 2008] Ameet Talwalkar, Sanjiv Kumar, and Henry A. Rowley. Large-scale manifold learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[Tukey, 1977] John W. Tukey. *Exploratory Data Analysis*. Pearson, 1977.

[Williams and Seeger, 2001] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, 2001.

[Williams *et al.*, 2002] Christopher Williams, Carl Edward Rasmussen, Anton Schwaighofer, and Volker Tresp. Observations on the nystrm method for gaussian processes. Technical report, Institute for Adaptive and Neural Computation, Division of Informatics, University of Edinburgh, 2002.

[Wu, 2010] Jianxin Wu. A fast dual method for HIK SVM learning. In *Proceedings of the 11th European Conference on Computer Vision*, pages 552–565, 2010.

[Zhang and Kwok, 2009] Kai Zhang and James T. Kwok. Density-weighted Nyström method for computing large kernel eigensystems. *Neural Computation*, 21(1):121–146, January 2009.

[Zhang *et al.*, 2008] Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1232–1239, 2008.