Improving the Modified Nyström Method Using Spectral Shifting

Shusen Wang[1] **Chao Zhang**[1] Hui Qian[1] Zhihua Zhang[2]

[1]Zhejiang University, China

[2]Shanghai Jiao Tong University, China

SIGKDD 2014

Outline

Outline

Kernel methods

- $\mathbf{K}$: $n \times n$ kernel matrix.
- Matrix inverse $\mathbf{b} = (\mathbf{K} + \alpha \mathbf{I}_n)^{-1} \mathbf{y}$
  - time complexity: $\mathcal{O}(n^3)$
  - performed by Gaussian process regression, least square SVM, kernel ridge regression
- Partial eigenvalue decomposition of $\mathbf{K}$
  - time complexity: $\mathcal{O}(n^2 k)$
  - performed by kernel PCA and some manifold learning methods

Kernel methods

- $\mathbf{K}$: $n \times n$ kernel matrix.
- Matrix inverse $\mathbf{b} = (\mathbf{K} + \alpha \mathbf{I}_n)^{-1}\mathbf{y}$
  - time complexity: $\mathcal{O}(n^3)$
  - performed by Gaussian process regression, least square SVM, kernel ridge regression
- Partial eigenvalue decomposition of $\mathbf{K}$
  - time complexity: $\mathcal{O}(n^2 k)$
  - performed by kernel PCA and some manifold learning methods

Kernel methods

- $\mathbf{K}$: $n \times n$ kernel matrix.
- Matrix inverse $\mathbf{b} = (\mathbf{K} + \alpha \mathbf{I}_n)^{-1}\mathbf{y}$
  - time complexity: $\mathcal{O}(n^3)$
  - performed by Gaussian process regression, least square SVM, kernel ridge regression
- Partial eigenvalue decomposition of $\mathbf{K}$
  - time complexity: $\mathcal{O}(n^2 k)$
  - performed by kernel PCA and some manifold learning methods

Computational Challenges

- High time complexities: $\mathcal{O}(n^3)$ or $\mathcal{O}(n^2 k)$
- High space complexity: $\mathcal{O}(n^2)$
  - the iterative algorithms go many passes through the data
  - very slow if the RAM is not large enough

Computational Challenges

- High time complexities: $\mathcal{O}(n^3)$ or $\mathcal{O}(n^2k)$
- High space complexity: $\mathcal{O}(n^2)$
  - the iterative algorithms go many passes through the data
  - very slow if the RAM is not large enough

How to Speedup

- If we can find a fast low-rank factorization

$$\underbrace{\mathbf{K}}_{n\times n} \approx \underbrace{\mathbf{D}}_{n\times d}\underbrace{\mathbf{D}^T}_{d\times n},$$

then $(\mathbf{K}+\alpha\mathbf{I}_n)^{-1}$ and the partial eigenvalue decomposition of $\mathbf{K}$ can be approximated solved highly efficiently.

How to Speedup: Example 1

- Suppose we have a low-rank factorization

$$\underbrace{\mathbf{K}}_{n\times n} \approx \underbrace{\mathbf{D}}_{n\times d}\underbrace{\mathbf{D}^T}_{d\times n}.$$

- Approximately compute the matrix inverse $(\mathbf{K} + \alpha \mathbf{I}_n)^{-1}$ as follows.
- Expand $(\mathbf{D}\mathbf{D}^T + \alpha \mathbf{I}_n)^{-1}$ using the Sherman-Morrison-Woodbury formula and obtain

$$\big(\mathbf{D}\mathbf{D}^T + \alpha \mathbf{I}_n\big)^{-1} = \alpha^{-1}\mathbf{I}_n - \alpha^{-1} \underbrace{\mathbf{D}}_{n\times d} \big(\underbrace{\alpha \mathbf{I}_d + \mathbf{D}^T\mathbf{D}}_{d\times d}\big)^{-1} \underbrace{\mathbf{D}^T}_{d\times n}.$$

It costs only $\mathcal{O}(nd^2)$ time and $\mathcal{O}(nd)$ space to compute

$$\mathbf{b} \; = \; \big(\mathbf{D}\mathbf{D}^T + \alpha \mathbf{I}_n\big)^{-1}\mathbf{y}.$$

How to Speedup: Example 2

- Suppose we have a low-rank factorization

$$\underbrace{\mathbf{K}}_{n\times n} \approx \underbrace{\mathbf{D}}_{n\times d} \underbrace{\mathbf{D}^T}_{d\times n},$$

- Compute the eigenvalue decomposition of **K** as follows.

- Compute the eigenvalue decomposition of the $d \times d$ small matrix $\mathbf{S} = \mathbf{D}^T\mathbf{D} \in \mathbb{R}^{d\times d}$:

$$\mathbf{S} \;=\; \mathbf{U_S\Lambda_S U_S^T}.$$

The partial eigenvalue decomposition of $\mathbf{DD}^T$ is

$$\mathbf{K} \;\approx\; \mathbf{DD}^T \;=\; \left(\mathbf{DU_S\Lambda_S}^{-1/2}\right) \mathbf{\Lambda_S} \left(\mathbf{DU_S\Lambda_S}^{-1/2}\right)^T$$

It costs only $\mathcal{O}(nd^2)$ time and $\mathcal{O}(nd)$ space.

Outline

The Nyström Method

- **Random Selection**:
  selects c ($\ll n$) columns of $\mathbf{K}$ to construct $\mathbf{C}$ using some randomized algorithms.
- **The Nyström Approximation: $\widetilde{\mathbf{K}} \approx \mathbf{K}$**

$$\underbrace{\widetilde{\mathbf{K}}}_{n\times n} = \underbrace{\mathbf{C}}_{n\times c}\underbrace{\mathbf{U}}_{c\times c}\underbrace{\mathbf{C}^T}_{c\times n}.$$

The Nyström Method

- **Random Selection**:
  selects c ($\ll n$) columns of $\mathbf{K}$ to construct $\mathbf{C}$ using some randomized algorithms.
- **The Nyström Approximation**: $\widetilde{\mathbf{K}} \approx \mathbf{K}$

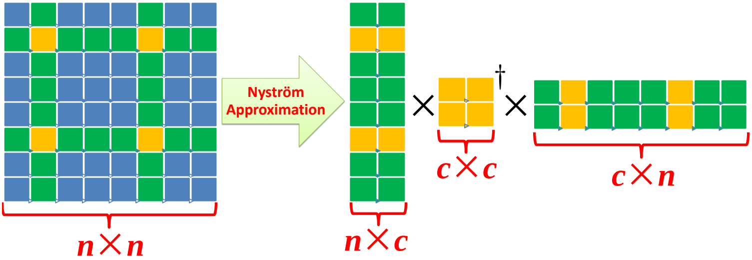
$$\underbrace{\widetilde{\mathbf{K}}}_{n\times n} = \underbrace{\mathbf{C}}_{n\times c}\underbrace{\mathbf{U}}_{c\times c}\underbrace{\mathbf{C}^T}_{c\times n}.$$

The Standard Nyström Approximation

- **The Standard Nyström Approximation**:

$$\mathbf{K} \quad \approx \quad \tilde{\mathbf{K}}_c^{\mathrm{nys}} \quad = \quad \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$$

(A low-rank factorization).



The Standard Nyström Approximation

Theorem (Gittens & Mahoney, ICML 2013)

Using a sampling algorithm, if $c = 3200\epsilon^{-1} k \log(16k/\delta)$ *the error incurred by the standard Nyström method satisfies*

$$\left\|\mathbf{K} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\right\|_F^2 \;\leq\; (1+\epsilon)\|\mathbf{K} - \mathbf{K}_k\|_F^2.$$

with probability at least $0.6 - \delta$.

The Modified Nyström Approximation

- **The Modified Nyström Approximation**:

$$\mathbf{K} \approx \tilde{\mathbf{K}}_c^{\mathrm{mod}} = \mathbf{C}\Big(\underbrace{\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^T}_{c\times c}\Big)\mathbf{C}^T$$

Theorem (Wang & Zhang, JMLR 2013)

Using a column sampling algorithm, the error incurred by the modified Nyström method satisfies

$$\mathbb{E}\big\|\mathbf{K} - \mathbf{C}\big(\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^T\big)\mathbf{C}^T\big\|_F^2 \leq \Big(1 + \sqrt{\frac{k}{c}}\Big)\|\mathbf{K} - \mathbf{K}_k\|_F^2.$$

The Modified Nyström Approximation

- **The Modified Nyström Approximation**:

$$\mathbf{K} \quad \approx \quad \tilde{\mathbf{K}}_c^{\mathrm{mod}} \quad = \quad \mathbf{C}\Big(\underbrace{\mathbf{C}^{\dagger}\mathbf{K}(\mathbf{C}^{\dagger})^{T}}_{c\times c}\Big)\mathbf{C}^{T}$$

Theorem (Wang & Zhang, JMLR 2013)

Using a column sampling algorithm, the error incurred by the modified Nyström method satisfies

$$\mathbb{E}\big\|\mathbf{K} - \mathbf{C}\big(\mathbf{C}^{\dagger}\mathbf{K}(\mathbf{C}^{\dagger})^{T}\big)\mathbf{C}^{T}\big\|_F^2 \;\leq\; \Big(1 + \sqrt{\frac{k}{c}}\Big)\|\mathbf{K} - \mathbf{K}_k\|_F^2.$$

Slowly Decayed Eigenvalues Spectrum Problem

- When the spectrum decays slowly,the approximation error could be rather large due to the large $\|\mathbf{K} - \mathbf{K}_k\|_F^2$.
- How to reduce $\|\mathbf{K} - \mathbf{K}_k\|_F^2$?

Outline

Modified Nyström by Spectral Shifting

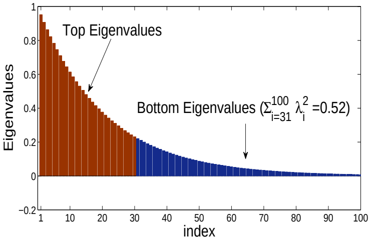
- Reduce $\|\mathbf{K} - \mathbf{K}_k\|_F^2$ by shifting the spectrum of $\mathbf{K}$ with $\mathbf{K} - \delta\mathbf{I}_n$, since $\lambda_i(\mathbf{K} - \delta\mathbf{I}_n) = \lambda_i(\mathbf{K}) - \delta$.
- The Modified Nyström by Spectral Shifting algorithm :
  - Shifting : give a shift to the original matrix $\bar{\mathbf{K}} = \mathbf{K} - \delta\mathbf{I}_n$
  - Approximation : approximate the matrix $\bar{\mathbf{K}} = \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T$
  - $\mathbf{K} \approx \tilde{\mathbf{K}}_c^{\mathrm{ss}} = \bar{\mathbf{K}} + \delta\mathbf{I}_m = \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta\mathbf{I}_n$

Modified Nyström by Spectral Shifting

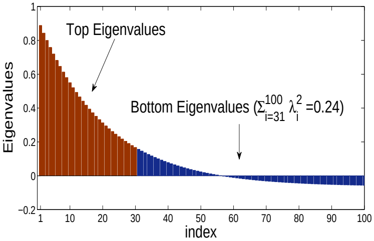
- Reduce $\|\mathbf{K} - \mathbf{K}_k\|_F^2$ by shifting the spectrum of $\mathbf{K}$ with $\mathbf{K} - \delta\mathbf{I}_n$, since $\lambda_i(\mathbf{K} - \delta\mathbf{I}_n) = \lambda_i(\mathbf{K}) - \delta$.
- The Modified Nyström by Spectral Shifting algorithm :
  - Shifting : give a shift to the original matrix $\bar{\mathbf{K}} = \mathbf{K} - \delta\mathbf{I}_n$
  - Approximation : approximate the matrix $\bar{\mathbf{K}} = \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T$
  - $\mathbf{K} \approx \tilde{\mathbf{K}}_c^{\mathrm{ss}} = \bar{\mathbf{K}} + \delta\mathbf{I}_m = \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta\mathbf{I}_n$

Eigenvalues Spectrum Shifting

Example 1: 100×100 SPSD matrix ,whose the t-th eigenvalue is 1.05^{-t}, $\|\mathbf{K} - \mathbf{K}_{30}\|_F^2$ before and after spectrum shifting.



(a) Before shift



(b) After shift

How to choose shift parameter δ

- It hold with probability that

$$\mathbb{E}\left\|\mathbf{K}-\tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\| = \mathbb{E}\left\|\bar{\mathbf{K}}-\bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T\right\| \leq (1+\epsilon)\left\|\bar{\mathbf{K}}-\bar{\mathbf{K}}_k\right\|_F.$$

- Try to minimize the following optimization problem to compute δ

$$\min_{\delta\geq 0} \ \left\|\bar{\mathbf{K}}-\bar{\mathbf{K}}_k\right\|_F^2; \qquad \text{s.t. } \bar{\mathbf{K}}=\mathbf{K}-\delta\mathbf{I}_n.$$

- Finally we can get

$$\mathbb{E}\left\|\mathbf{K}-\tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\|_F^2 \leq (1+\epsilon)\left(\left\|\mathbf{K}-\mathbf{K}_k\right\|_F^2 - \frac{\left[\sum_{i=k+1}^{n}\lambda_i(\mathbf{K})\right]^2}{n-k}\right).$$

How to choose shift parameter δ

- It hold with probability that

$$\mathbb{E}\left\|\mathbf{K}-\tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\| = \mathbb{E}\left\|\bar{\mathbf{K}}-\bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T\right\| \leq (1+\epsilon)\left\|\bar{\mathbf{K}}-\bar{\mathbf{K}}_k\right\|_F.$$

- Try to minimize the following optimization problem to compute δ

$$\min_{\delta\geq 0} \left\|\bar{\mathbf{K}}-\bar{\mathbf{K}}_k\right\|_F^2; \qquad \text{s.t. } \bar{\mathbf{K}} = \mathbf{K}-\delta\mathbf{I}_n.$$

- Finally we can get

$$\mathbb{E}\left\|\mathbf{K}-\tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\|_F^2 \leq (1+\epsilon)\left(\left\|\mathbf{K}-\mathbf{K}_k\right\|_F^2 - \frac{\left[\sum_{i=k+1}^n \lambda_i(\mathbf{K})\right]^2}{n-k}\right).$$

How to choose shift parameter δ

- It hold with probability that

$$\mathbb{E}\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\| = \mathbb{E}\left\|\bar{\mathbf{K}} - \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T\right\| \leq (1+\epsilon)\left\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\right\|_F.$$

- Try to minimize the following optimization problem to compute δ

$$\min_{\delta \geq 0} \left\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\right\|_F^2; \qquad \text{s.t. } \bar{\mathbf{K}} = \mathbf{K} - \delta \mathbf{I}_n.$$

- Finally we can get

$$\mathbb{E}\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\|_F^2 \leq (1+\epsilon)\left(\|\mathbf{K} - \mathbf{K}_k\|_F^2 - \frac{\left[\sum_{i=k+1}^n \lambda_i(\mathbf{K})\right]^2}{n-k}\right).$$

Better Upper Bound

- $\delta^{\mathrm{opt}} = \frac{1}{n-k}\left(\mathrm{tr}(\mathbf{K}) - \sum_{j=1}^{k} \sigma_j(\mathbf{K})\right)$.
- Random projection can be used to speedup the computation of top-k SVD.

Theorem

Using a column sampling algorithm, the error incurred by the modified Nyström by Spectral Shifting method satisfies

$$\mathbb{E}\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{ss}\right\|_F^2 \leq \left(1 + \sqrt{\frac{c}{k}}\right)\left(\|\mathbf{K} - \mathbf{K}_k\|_F^2 - \frac{\left[\sum_{i=k+1}^{n} \lambda_i(\mathbf{K})\right]^2}{n-k}\right).$$

Better Upper Bound

- **Example 2:** Let $\mathbf{K}$ be an $m \times m$ SPSD matrix such that $\lambda_1(\mathbf{K}) \geq \cdots \geq \lambda_k(\mathbf{K}) > \theta = \lambda_{k+1}(\mathbf{K}) = \cdots = \lambda_m(\mathbf{K}) > 0$. The approximation error of the three different Nyström method is :

$$\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\mathrm{ss}}\right\|_F^2 = 0,$$

and that

$$\begin{aligned}(m-c)\theta^2 &= \left\|\mathbf{K} - \mathbf{K}_c\right\|_F^2 \\ &\leq \left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\mathrm{mod}}\right\|_F^2 \leq \left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\mathrm{nys}}\right\|_F^2.\end{aligned}$$

Application to Kernel Method

- Use SS-Nyström $\mathbf{K} \approx \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta \mathbf{I}_n$ to speedup matrix inverse and eigenvalue decomposition
- In matrix inverse $(\mathbf{K} + \alpha \mathbf{I}_n)^{-1}$,

$$(\tilde{\mathbf{K}}_c^{\mathrm{ss}} + \alpha \mathbf{I}_m)^{-1} = \tau^{-1}\mathbf{I}_m - \tau^{-1}\bar{\mathbf{C}}\mathbf{Z}(\tau \mathbf{\Lambda}^{-1} + \mathbf{Z}^T\bar{\mathbf{C}}^T\bar{\mathbf{C}}\mathbf{Z})^{-1}\mathbf{Z}^T\bar{\mathbf{C}}^T,$$

where $\tau = \delta + \alpha$, $\bar{\mathbf{U}} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T$, $\rho = \mathrm{rank}(\bar{\mathbf{U}})$.

- In eigenvalue decomposition of K,

$$\tilde{\mathbf{K}}_c^{\mathrm{ss}} = (\mathbf{U_C U_S})(\mathbf{\Lambda_S} + \delta \mathbf{I}_\rho)(\mathbf{U_C U_S})^T + \mathbf{U}_\perp(\delta \mathbf{I}_n)\mathbf{U}_\perp^T,$$

where $\mathbf{S} = \mathbf{\Sigma}_{\bar{\mathbf{C}}}\mathbf{V}_{\bar{\mathbf{C}}}\bar{\mathbf{U}}\mathbf{V}_{\bar{\mathbf{C}}}^T\mathbf{\Sigma}_{\bar{\mathbf{C}}}^T$, $\rho = \mathrm{rank}(\bar{\mathbf{C}})$.

Application to Kernel Method

- Use SS-Nyström $\mathbf{K} \approx \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta \mathbf{I}_n$ to speedup matrix inverse and eigenvalue decomposition
- In matrix inverse $(\mathbf{K} + \alpha \mathbf{I}_n)^{-1}$,

$$\left(\tilde{\mathbf{K}}_c^{\mathrm{ss}} + \alpha \mathbf{I}_m\right)^{-1} = \tau^{-1}\mathbf{I}_m - \tau^{-1}\bar{\mathbf{C}}\mathbf{Z}\left(\tau \mathbf{\Lambda}^{-1} + \mathbf{Z}^T\bar{\mathbf{C}}^T\bar{\mathbf{C}}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\bar{\mathbf{C}}^T,$$

where $\tau = \delta + \alpha$, $\bar{\mathbf{U}} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T$, $\rho = \mathrm{rank}(\bar{\mathbf{U}})$.

- In eigenvalue decomposition of K,

$$\tilde{\mathbf{K}}_c^{\mathrm{ss}} = (\mathbf{U}_\mathbf{C}\mathbf{U}_\mathbf{S})(\mathbf{\Lambda}_\mathbf{S} + \delta \mathbf{I}_\rho)(\mathbf{U}_\mathbf{C}\mathbf{U}_\mathbf{S})^T + \mathbf{U}_\perp(\delta \mathbf{I}_n)\mathbf{U}_\perp^T,$$

where $\mathbf{S} = \mathbf{\Sigma}_{\bar{\mathbf{C}}}\mathbf{V}_{\bar{\mathbf{C}}}\bar{\mathbf{U}}\mathbf{V}_{\bar{\mathbf{C}}}^T\mathbf{\Sigma}_{\bar{\mathbf{C}}}^T$, $\rho = \mathrm{rank}(\bar{\mathbf{C}})$.

Application to Kernel Method

- Use SS-Nyström $\mathbf{K} \approx \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta\mathbf{I}_n$ to speedup matrix inverse and eigenvalue decomposition
- In matrix inverse $(\mathbf{K} + \alpha\mathbf{I}_n)^{-1}$,

$$\left(\tilde{\mathbf{K}}_c^{\mathrm{ss}} + \alpha\mathbf{I}_m\right)^{-1} = \tau^{-1}\mathbf{I}_m - \tau^{-1}\bar{\mathbf{C}}\mathbf{Z}\left(\tau\mathbf{\Lambda}^{-1} + \mathbf{Z}^T\bar{\mathbf{C}}^T\bar{\mathbf{C}}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\bar{\mathbf{C}}^T,$$

where $\tau = \delta + \alpha$, $\bar{\mathbf{U}} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T$, $\rho = \mathrm{rank}(\bar{\mathbf{U}})$.

- In eigenvalue decomposition of K,

$$\tilde{\mathbf{K}}_c^{\mathrm{ss}} = (\mathbf{U_C}\mathbf{U_S})(\mathbf{\Lambda_S} + \delta\mathbf{I}_\rho)(\mathbf{U_C}\mathbf{U_S})^T + \mathbf{U}_\perp(\delta\mathbf{I}_n)\mathbf{U}_\perp^T,$$

where $\mathbf{S} = \mathbf{\Sigma}_{\bar{\mathbf{C}}}\mathbf{V}_{\bar{\mathbf{C}}}\bar{\mathbf{U}}\mathbf{V}_{\bar{\mathbf{C}}}^T\mathbf{\Sigma}_{\bar{\mathbf{C}}}^T$, $\rho = \mathrm{rank}(\bar{\mathbf{C}})$.

Outline

Experiment setup

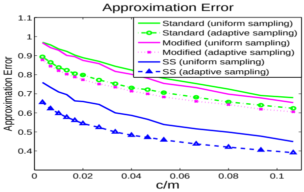
- Tested on datasets released by UCI[Frank & Asuncion,2010]
- A RBF kernel $\mathbf{K}_\alpha$ is defined by

$$k_{ij}^{\alpha} = \exp\left(- \frac{1}{2\alpha} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \right)$$

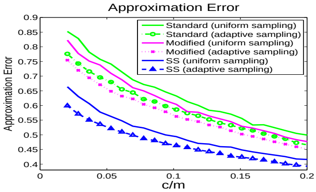
- Two sampling method:
  - uniform sampling[Gittens & Mahoney 2013]
  - adaptive sampling[Wang & Zhang,2013]

Kernel approximation error of different method

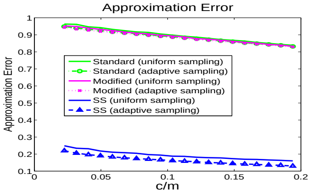
Kernel approximation error on different real world RBF kernels($\alpha = 0.2$)



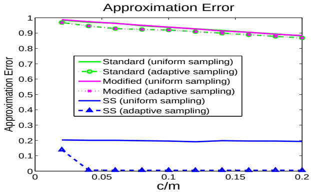
(a) Letters



(b) Wine Quality



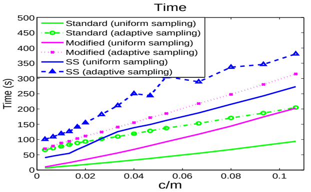
(c) Satimage



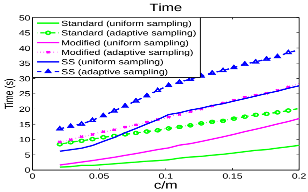
(d) Splice

Kernel approximation elapsed time of different method

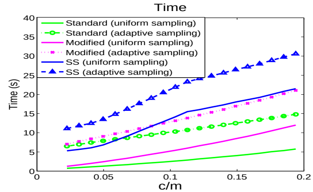
Kernel approximation elapsed time on different real world RBF kernels($\alpha = 0.2$)



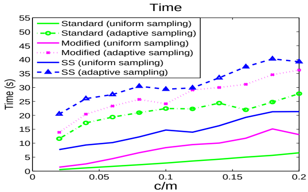
(a) Letters



(b) Wine Quality



(c) Satimage



(d) Splice

Reference

S. Wang and Z. Zhang: *Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling.* JMLR, 2013.

N. Halko, P.-G. Martinsson, and J. A. Tropp: *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.* SIAM Review, 53(2):217–288, 2011.

A. Frank and A. Asuncion: UCI machine learning repository, 2010.

A. Gittens and M. W. Mahoney: Revisiting the nyström method for improved large-scale machine learning. International Conference on Machine Learning (ICML), 2013.