Improved Nyström Low Rank Approximation and Error Analysis

Kai Zhang Ivor W. Tsang James T. Kwok

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Outline

Outline

Low-rank Approximation of Kernel Matrices

Definition

Given $n \times n$ kernel matrix K (or its variants)

- find $K \approx GG'$
- $G \in \mathbb{R}^{n \times m}$, $m \ll n$
- measured by $\|K - GG'\|_F$

Optimal solution

- eigenvalue decomposition
- $O(n^3)$, too expensive

Efficient alternatives needed

Applications

Applications

1. Matrix Eigenvalue Decomposition

If $K \approx GG'$, then top m eigen-vectors/values (U, Σ) of K can be approximated by

$$U \approx GV\Lambda^{-\frac{1}{2}}, \Sigma \approx \Lambda,$$

where $G'G = V\Lambda V'$.

2. Matrix Inverse, Linear System

Woodbury Formula

$$(K + \sigma I)^{-1} \simeq \frac{1}{\sigma}\left(I - G(\sigma I + G'G)^{-1}G'\right)$$

Examples

Algorithms that can benefit from low-rank approximation

- Kernel PCA [Schölkopf et. al. 1998]
- Kernel LDA [Mika et. al. 1999]
- Spectral Clustering [Ng and Jordan 2001]
- Laplacian Eigenmap [Belkin and Niyogi 2002]
- Normalized-cut [Shi and Malik 2000]
- Gaussian Process [Williams and Seeger 2001]
- Least square SVM [Suykens and Vandewalle 1999]

Scale-Up Methods

Greedy Approaches $O(m^2 n)$

- Incomplete Cholesky decomposition
  [Bach & Jordan, 2002; Fine & Scheinberg, 2001]
- Sparse greedy kernel methods [Smola & Schökopf, 2000]
- Greedy spectral embedding [Ouimet & Bengio 2005]

Random Sampling $O(m^3 + mn)$

- Nyström Method
  [Williams & Seeger, 2001; Lawrence & Herbrich, 2005]

Randomized Algorithms $O(n^2)$

- Probabilistic sampling [Drineas & Mahoney 2005]

The Nyström Method

Given n samples: $\mathcal{X}$, solve integral equation

$$\int p(y)k(x,y)\phi(y)dy = \lambda\phi(x)$$

1. Empirical approximation

$$\frac{1}{m}\sum_{j=1}^{m} k(x,z_j)\phi(z_j) = \lambda\phi(x).$$

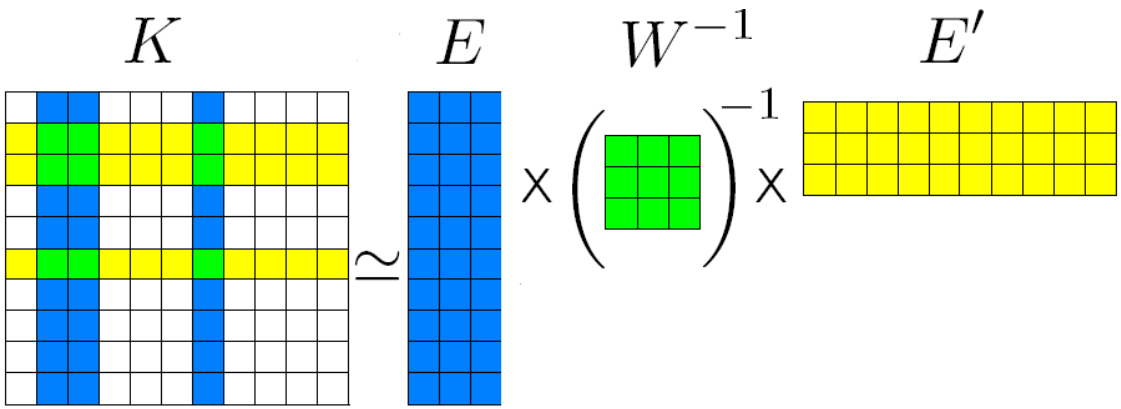
   - m landmark points: $\mathcal{Z}$

2. Small eigen-system $W\phi_Z = m\lambda_Z\phi_Z$
   - W: $m \times m$ kernel matrix on $\mathcal{Z}$

3. Extrapolate $\phi_X = (m\lambda_Z)^{-1}E\phi_Z$.
   - E: $n \times m$ kernel matrix on $(\mathcal{X}, \mathcal{Z})$

Nyström: Matrix Completion View

1. Kernel matrix $K_{n\times n}$ on $\mathcal{X}$;
2. Randomly choose $m \ll n$ columns $E_{n\times m}$
   - corresponds to landmark set $\mathcal{Z}$, $|\mathcal{Z}|=m$
   - $W_{m\times m}$: kernel matrix on $\mathcal{Z}$
3. Reconstruct by $K \approx EW^{-1}E'$



Outline

Approximation Error

Nyström low rank approximation error

$$\mathcal{E} = \|K - EW^{-1}E'\|_F$$

Error analysis difficult

- K, W and E are all of different sizes
- no clear data structure to utilize

A probabilistic bound

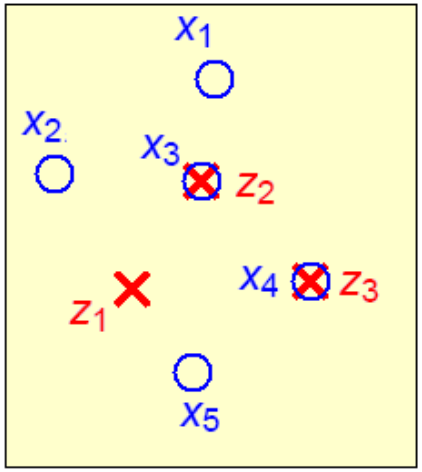
- probabilistic sampling [Drineas 2005]
- sampling probabilities expensive to compute
- worse than random sampling

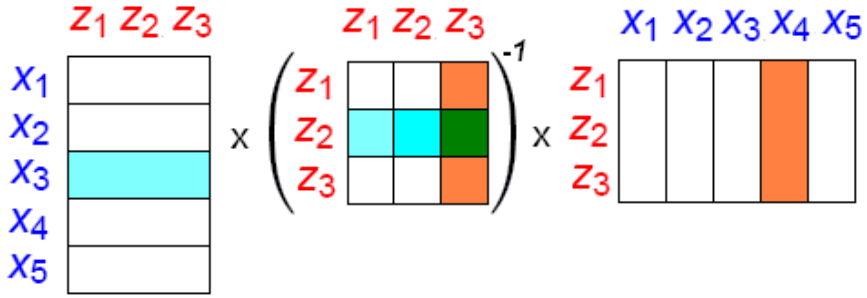
Important Observation

Proposition

Given data set $\mathcal{X} = \{x_i\}_{i=1}^n$, *landmark set* $\mathcal{Z} = \{z_j\}_{j=1}^m$, *the Nyström method can accurately reconstruct* $K(x_i, x_j)$ *if*

$$\exists\, p, q, \text{ such that } z_p = x_i, \;\; z_q = x_j .$$





$$WW^{-1}W = W$$

Indication

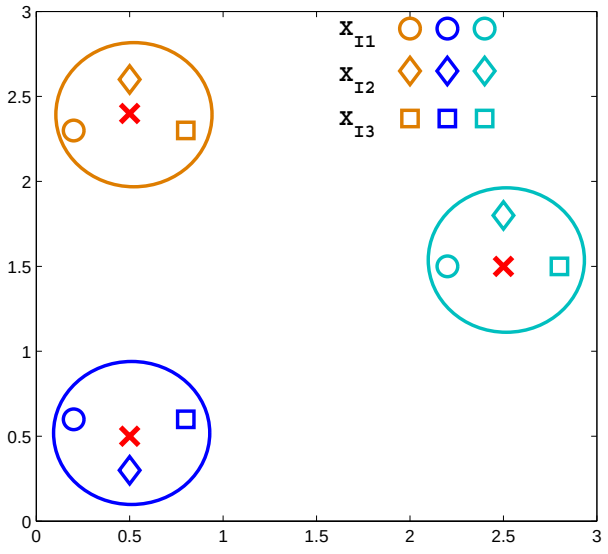
Intuitive Criteria

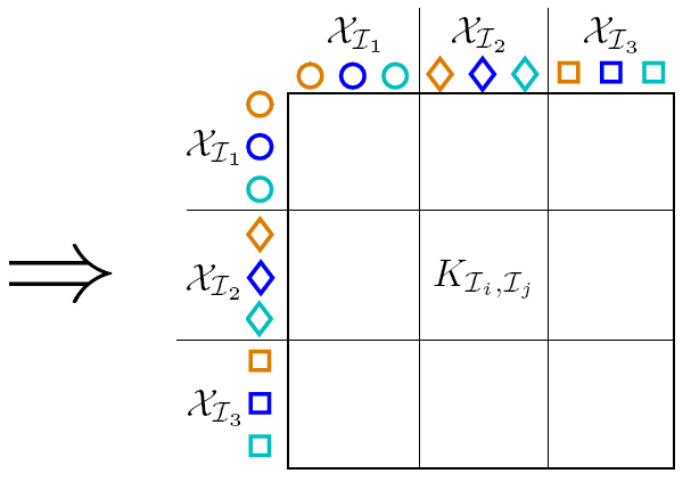
Find $\mathcal{Z}$ to "*overlap*" with $\mathcal{X}$ as much as possible

Suppose $\mathcal{X}$ aggregates into m clusters

- Ideal case: each cluster shrinks to a point
  - optimal strategy: choosing cluster centers as $\mathcal{Z}$
  - zero approximation error
- Practical case: each cluster is diffused
  - still cluster centers?
  - how to cluster?

Partial Approximation Error





1. Partition $\mathcal{X}$ into m clusters S_i's by "seeds" $\mathcal{Z}$
2. For $t = 1, 2, ..., T_{=\max|S_i|}$, sample $\mathcal{X}_{\mathcal{I}_t} = \{x_j \in S_i\}_{i=1}^{m}$:
   $\mathcal{X} = \mathcal{X}_{\mathcal{I}_1} \cup \mathcal{X}_{\mathcal{I}_1} \cup ... \cup \mathcal{X}_{\mathcal{I}_T}$
3. Define partial approximation error $\mathcal{E}_{\mathcal{I}_i,\mathcal{I}_j} = \|K_{\mathcal{I}_i,\mathcal{I}_j} - E_{\mathcal{I}_i,\mathcal{Z}} W^{-1} E'_{\mathcal{I}_j,\mathcal{Z}}\|_F$

Partial Approximation Error

The *partial approximation error* $\mathcal{E}_{\mathcal{I}_i,\mathcal{I}_j}$ is bounded by

$$\sqrt{2mC_{\mathcal{X}}^{k}(e_{\mathcal{I}_i}+e_{\mathcal{I}_j})}+\sqrt{mC_{\mathcal{X}}^{k}e_{\mathcal{I}_i}}$$
$$+\sqrt{mC_{\mathcal{X}}^{k}e_{\mathcal{I}_j}}+mC_{\mathcal{X}}^{k}\sqrt{e_{\mathcal{I}_i}e_{\mathcal{I}_j}}\|W^{-1}\|_F.$$

- $K_{\mathcal{I}_i,\mathcal{I}_j}$: similarity matrix on $(\mathcal{X}_{\mathcal{I}_i},\mathcal{X}_{\mathcal{I}_j})$
- $E_{\mathcal{I}_i,\mathcal{Z}}$: similarity matrix on $(\mathcal{X}_{\mathcal{I}_i},\mathcal{Z})$
- W: kernel matrix on $(\mathcal{Z},\mathcal{Z})$
- $C_{\mathcal{X}}^{K}$: kernel dependent constant
- $e_{\mathcal{I}_i}$: partial quantization error $\sum_{x_i\in\mathcal{X}_{\mathcal{I}_i}}\left\|x_i-z_{c(i)}\right\|^2$

Complete Approximation Error

Proposition

The complete approximation error $\mathcal{E}$ is bounded by

$$\mathcal{E} \leq 4T\sqrt{mC_{\mathcal{X}}^{k}eT} + mC_{\mathcal{X}}^{k}Te\|W^{-1}\|_{F}$$

- $e = \sum_{i=1}^{n} \left\|x_i - z_{c(i)}\right\|^2$
  - *total quantization error of coding $x_i \in \mathcal{X}$ with closest $z_j \in \mathcal{Z}$*
- *applicable kernels*
  - *linear, polynomial, RBF (Gaussian, Laplacian)*

New Sampling Scheme

Key factor of Nyström low-rank approximation

- encoding power of landmark set $\mathcal{Z}$

k-means

- find local optimum of quantization error

Use k-means centers as landmark points

- linear time & space complexity
- fast algorithms available
  [Pelleg & Moore 1999], [Kanungo et al., 2001], [Elkan, 2003]
- empirically outperforms all Nyström variants

Illustration

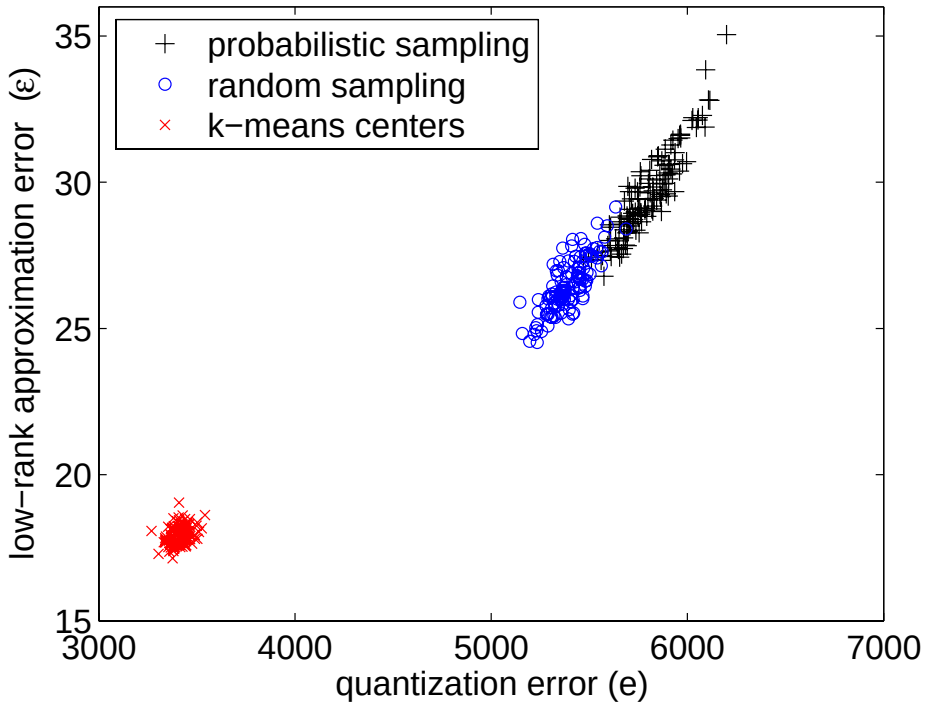


Figure: Low rank approximation error $\mathcal{E}$ is strongly correlated with quantization error e.

Outline

Kernel Principal Component Analysis

Experimental setting

- Data sets
  - LIBSVM (UCI), MNIST, USPS
- Tasks
  - Matrix low-rank approximation
  - kernel PCA
  - Least square SVM
- Comparisons
  - random sampling
  - probabilistic sampling
  - Incomplete Cholesky decomposition
  - eigenvalue decomposition (SVD, optimal)

Experiments: Low Rank Approximation

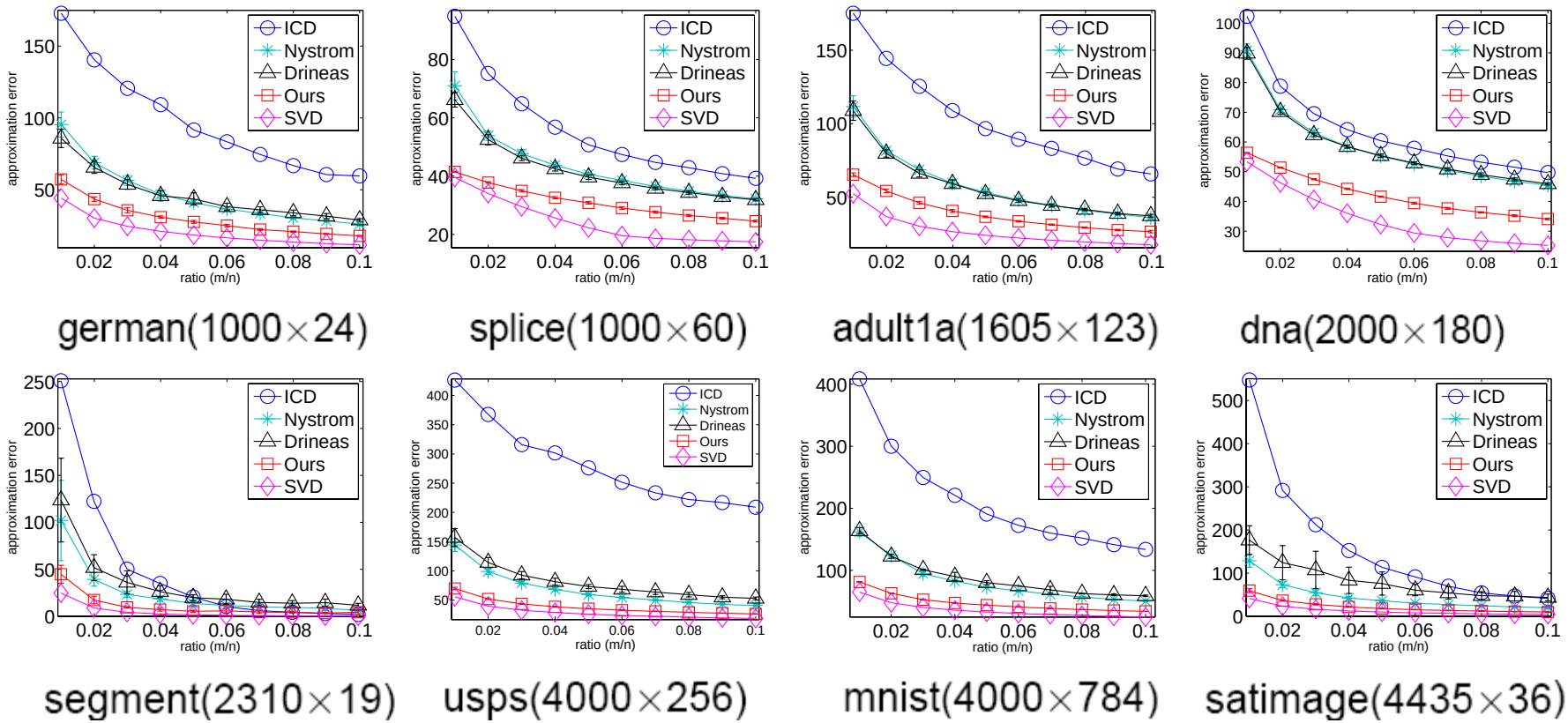


Figure: Low rank approximation error by SVD, ICD, random, probabilistic, and our sampling scheme (RBF kernel).

Experiments: Low Rank Approximation

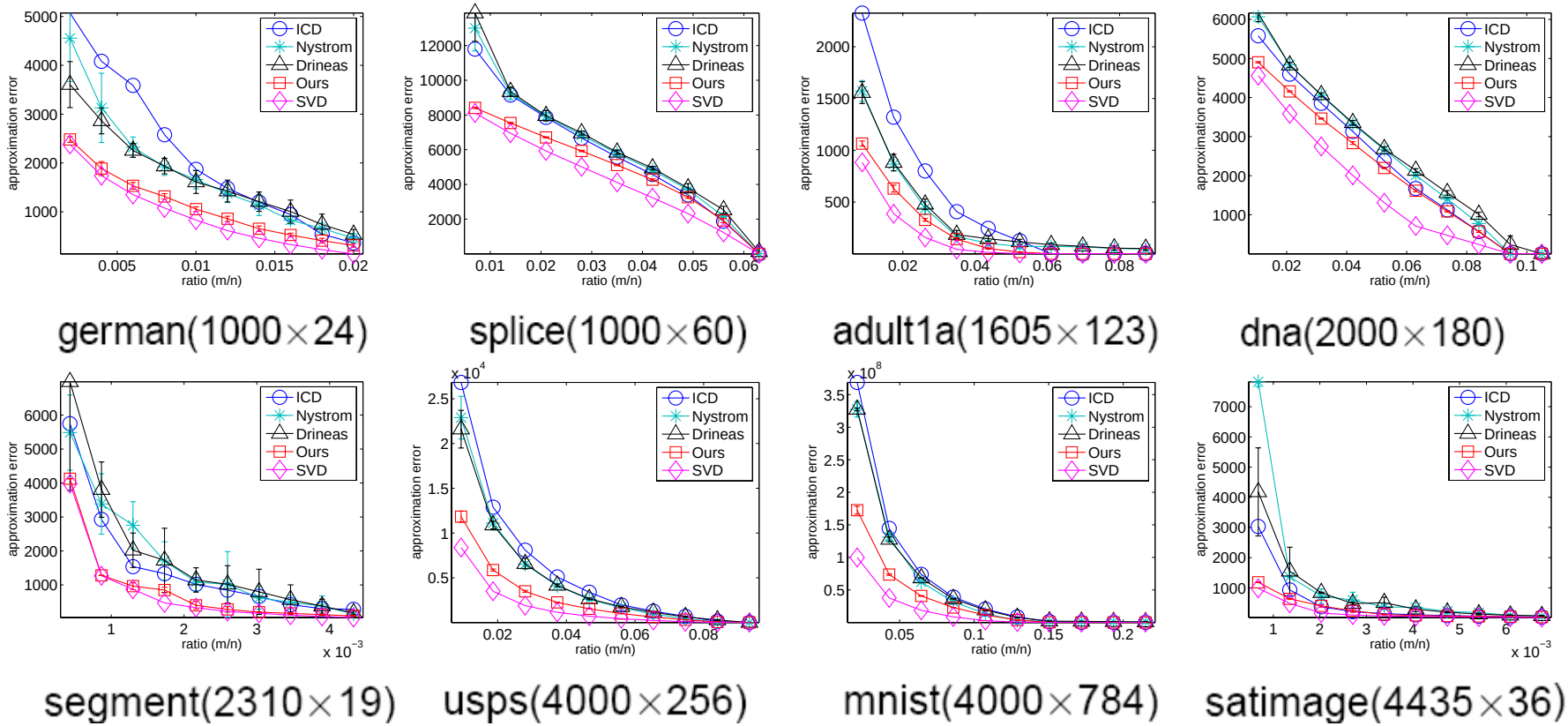


Figure: Low rank approximation error by SVD, ICD, random, probabilistic, and our sampling scheme (linear kernel).

Experiments: Low Rank Approximation

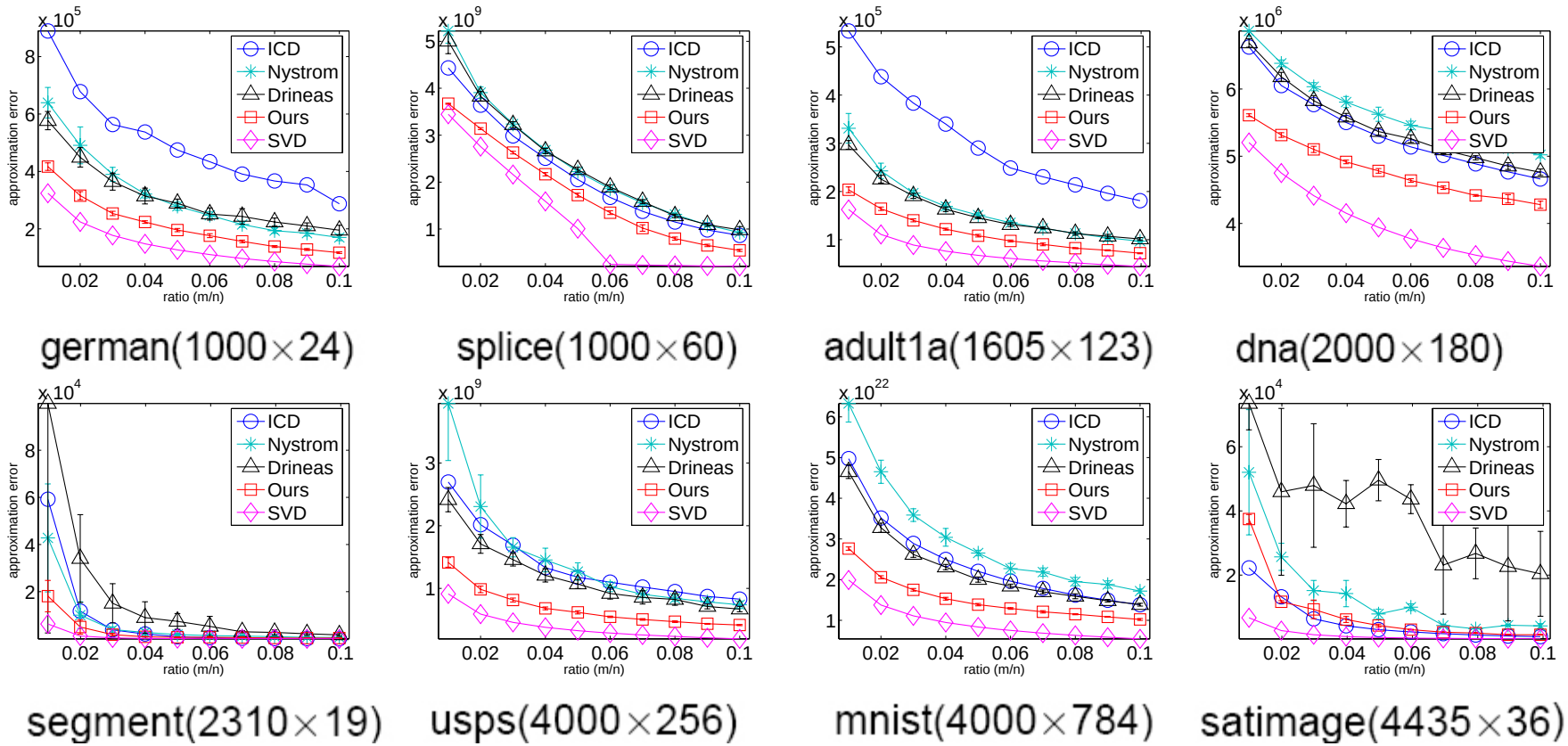


Figure: Low rank approximation error by SVD, ICD, random, probabilistic, and our sampling scheme (polynomial kernel).

Kernel Principal Component Analysis

Embedding errors (top3PC) and CPU time of low-rank-approx schemes.

	approximation error			
data	Ours	Nyström	Drineas	ICD
german	$\mathbf{(4.4\pm0.5)\times10^{-2}}$	$(2.6\pm0.5)\times10^{-1}$	$(2.7\pm0.3)\times10^{-1}$	5.1×10^{-1}
splice	$\mathbf{(3.4\pm0.4)\times10^{-1}}$	$(1.0\pm0.1)\times10^{0}$	$(1.1\pm0.1)\times10^{0}$	1.3×10^{0}
adt1a	$\mathbf{(4.4\pm0.4)\times10^{-2}}$	$(2.8\pm0.4)\times10^{-1}$	$(2.8\pm0.6)\times10^{-1}$	6.2×10^{-1}
dna	$\mathbf{(1.8\pm0.2)\times10^{-1}}$	$(1.1\pm0.1)\times10^{0}$	$(1.0\pm0.1)\times10^{0}$	1.2×10^{0}
sgment	$\mathbf{(7.8\pm4.4)\times10^{-4}}$	$(8.3\pm4.1)\times10^{-3}$	$(1.4\pm1.0)\times10^{-2}$	2.4×10^{-2}
w1a	$\mathbf{(1.5\pm0.7)\times10^{-1}}$	$(2.8\pm0.6)\times10^{-1}$	$(6.0\pm3.4)\times10^{-1}$	1.1×10^{0}
usps	$\mathbf{(1.1\pm0.1)\times10^{-2}}$	$(1.2\pm0.3)\times10^{-1}$	$(1.5\pm0.4)\times10^{-1}$	6.6×10^{-1}
stimage	$\mathbf{(5.2\pm0.9)\times10^{-4}}$	$(6.2\pm0.3)\times10^{-3}$	$(6.8\pm1.0)\times10^{-2}$	2.5×10^{-2}
SVD (time)	time consumption (seconds)			
german (27)	0.8	**0.03**	0.3	0.09
splice (24)	0.9	**0.05**	0.6	0.1
adt1a (134)	3.0	**0.2**	4.0	0.7
dna (197)	6.6	**0.5**	10.6	1.5
sgment (322)	4.2	**0.3**	1.8	1.0
w1a (394)	12.8	**1.8**	35.3	3.6
usps (736)	24.88	**1.86**	42.04	5.11
stimage (2762)	16.1	**1.5**	15.9	7.5

Experiments: Least Square SVM

Testing errors (in %) on different pairs of the USPS digits.

			$m = 0.05n$			
data	LS-SVM	SVD	Ours	Nyström	Drineas	ICD
1-7	0.97	0.97	0.77±0.09	1.56±0.51	2.69±1.64	4.62
2-3	2.47	2.47	2.91±0.55	4.48±1.51	4.54±1.16	4.12
2-5	1.67	0.83	1.91±0.62	2.82±1.10	2.98±0.73	5.30
3-8	2.71	3.31	3.93±0.84	5.64±1.45	5.52±1.21	6.02
5-8	2.76	2.14	3.38±0.65	4.34±1.52	4.28±1.38	4.29
6-8	0.59	0.59	1.04±0.35	2.69±1.34	2.24±0.98	5.65
8-9	1.45	1.16	1.04±0.50	2.79±2.00	2.52±0.85	2.33
2-7	1.44	0.86	0.90±0.10	1.47±0.64	2.45±1.25	3.76
3-5	4.91	4.91	6.53±1.18	7.39±1.39	7.03±1.25	7.36
4-7	2.88	2.59	2.54±0.40	3.28±0.96	3.30±1.03	8.06
5-6	1.21	0.60	1.57±0.63	2.79±1.25	2.52±1.33	3.93

From best to worst: red, green, blue, black.

Experiments: Least Square SVM

Testing errors (in %) on different pairs of the USPS digits.

	$m = 0.1n$					
data	LS-SVM	SVD	Ours	Nyström	Drineas	ICD
1-7	0.97	0.97	0.81±0.11	0.94±0.18	1.54±0.35	1.22
2-3	2.47	2.47	2.72±0.22	2.77±0.59	2.97±0.59	4.94
2-5	1.67	1.67	1.29±0.31	1.57±0.46	1.58±0.51	2.23
3-8	2.71	2.71	2.70±0.23	3.78± 0.53	3.97±0.55	4.21
5-8	2.76	2.14	2.66±0.34	2.82±0.53	3.12±0.58	4.60
6-8	0.59	0.89	0.61±0.25	1.03±0.45	1.18±0.44	5.05
8-9	1.45	1.45	1.13±0.23	1.22±0.51	1.79±0.37	2.91
2-7	1.44	0.86	0.96±0.17	0.87±0.09	1.08±0.39	2.61
3-5	4.91	3.98	5.25±0.64	5.22±0.77	5.49±0.83	5.52
4-7	2.88	2.59	2.60±0.26	2.41±0.36	2.43±0.45	5.47
5-6	1.21	1.21	1.27±0.32	1.25±0.41	1.48±0.29	2.12

From best to worst: red, green, blue, black.

Outline

Conclusions

- Conclusion
  - Nyström low rank approximation error
    - depends crucially on encoding powers of landmark points
  - k-means is a very suitable sampling scheme
- Future work
  - different kernels may favor different sampling
  - extend to more general kernel matrices, and SVD
  - combine with side information (say, label info.)

Q & A

Thank you!