ICML@NYC
International Conference on Machine Learning

The Variational Nyström Method for Large-Scale Spectral Problems

Max Vladymyrov
Google Inc.

Miguel Carreira-Perpiñán
EECS, UC Merced

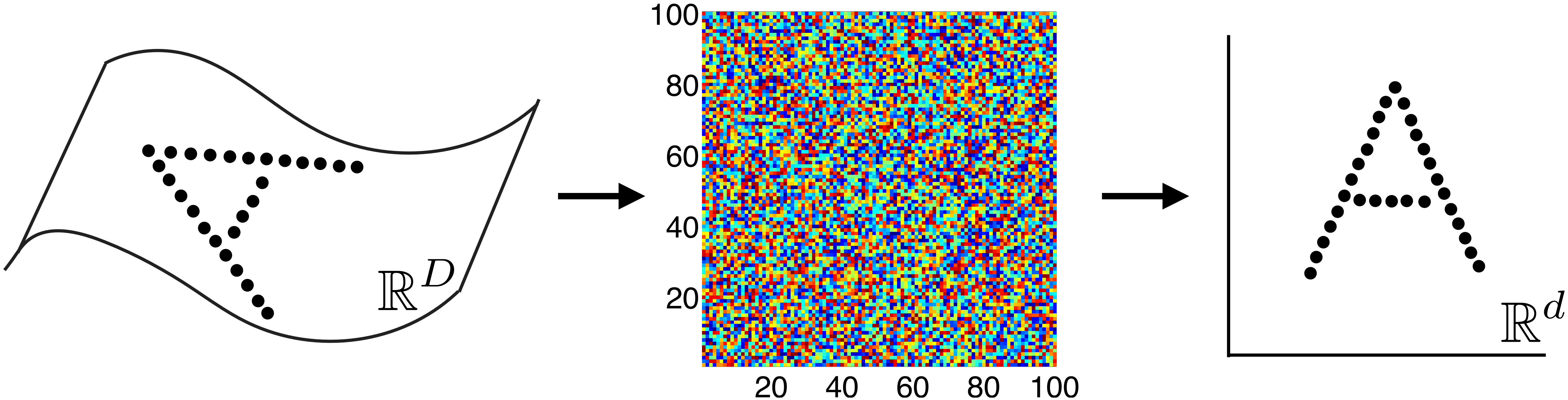
June 20, 2016

Graph based dimensionality reduction methods

Given high-dimensional data points $\mathbf{Y}_{D \times N} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$.

1. Convert data points to a $N \times N$ *affinity* matrix $\mathbf{M}$.
2. Find low-dimensional coordinates $\mathbf{X}_{d \times N} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, so that their similarity is as close as possible to $\mathbf{M}$.

High-dimensional input $\mathbf{Y}$ | Affinity $\mathbf{M}$ | Low-dimensional output $\mathbf{X}$



Spectral methods

- Consider a spectral problem:

$$\min_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}\mathbf{M}\mathbf{X}^T\right) \quad \text{s.t.} \quad \mathbf{X}\mathbf{X}^T = \mathbf{I},$$

  - $\mathbf{M}_{N\times N}$: symmetric psd affinity matrix.

- Examples:
  - Laplacian eigenmaps, $\mathbf{M}$ is a graph Laplacian.
  - ISOMAP, $\mathbf{M}$ is given by a matrix of shortest distances.
  - Kernel PCA, MDS, Locally Linear Embedding (LLE), etc.

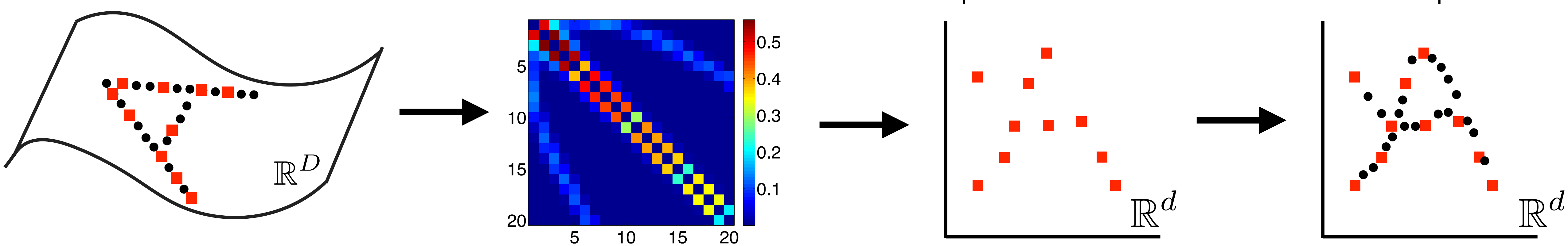
- Solution is unique and can be found in closed form from the eigenvectors of $\mathbf{M}$: $\mathbf{X} = \mathbf{U}_{\mathbf{M}}^T$.

With large N, solving the eigenproblem is infeasible even if $\mathbf{M}$ is sparse.

Learning with landmarks

Goal is find a fast, approximate solution for the embedding $\mathbf{X}$ using only the subset of the original points from $\mathbf{Y}$.

Select L landmarks (e.g. random subset) | Compute reduced $L \times L$ affinity matrix | Learn landmark representation | Project the rest of the points



Nyström method

Writing the affinity matrix $\mathbf{M}$ by blocks (landmarks first):

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B}_{21} \end{pmatrix}$$

The approximation to the eigendecomposition is equal to:

$$\widetilde{\mathbf{U}}_{\mathbf{M}} = \begin{pmatrix} \mathbf{U}_{\mathbf{A}} \\ \mathbf{B}_{21}\mathbf{U}_{\mathbf{A}}\boldsymbol{\Lambda}_{\mathbf{A}}^{-1} \end{pmatrix}$$

Essentially, an out-of-sample formula:

1. Solve the eigenproblem for a subset of points.
2. Predict the rest of the points through the interpolation formula.

Column Sampling method

Writing the affinity matrix $\mathbf{M}$ by blocks (landmarks first):

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B}_{21} \end{bmatrix}$$

The approximation to the eigendecomposition is given by the left singular vectors of $\mathbf{C}$:

$$\mathbf{C} = \mathbf{U_C}\boldsymbol{\Sigma}_\mathbf{C}\mathbf{V}_\mathbf{C}^T \quad \Rightarrow \quad \widetilde{\mathbf{U}}_\mathbf{M} = \mathbf{U_C}$$

Uses more information from the affinity matrix $\mathbf{M}$ than Nyström, but still ignores non-landmark/non-landmark interaction part $\mathbf{B}_{22}$.

Locally Linear Landmarks (LLL) (Vladymyrov & Carreira-Perpiñán, 2013)

- Construct the local linear projection matrix $\mathbf{Z}$ from the input $\mathbf{Y}$:

$$\mathbf{y}_n \approx \sum_{l=1}^{L} z_{ln}\widetilde{\mathbf{y}}_l, n = 1, \dots, N \quad \Rightarrow \quad \mathbf{Y} \approx \widetilde{\mathbf{Y}}\mathbf{Z}^T$$

- Additional assumption: this projection is satisfied in the embedding space: $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}^T$.
- Plugging the projection to the original obj. function:

$$\min_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}\mathbf{M}\mathbf{X}^T\right) \quad \text{s.t.} \quad \mathbf{X}\mathbf{X}^T = \mathbf{I},\ \mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}^T$$

$$\Downarrow$$

$$\min_{\widetilde{\mathbf{X}}} \operatorname{tr}\left(\widetilde{\mathbf{X}}\mathbf{Z}^T\mathbf{M}\mathbf{Z}\widetilde{\mathbf{X}}^T\right) \quad \text{s.t.} \quad \widetilde{\mathbf{X}}\mathbf{Z}^T\mathbf{Z}\widetilde{\mathbf{X}}^T = \mathbf{I}$$

- The solution is given by the reduced generalized eigenproblem:

$$\widetilde{\mathbf{X}} = \operatorname{eig}(\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}\mathbf{Z}^T)$$

- Final embedding are predicted as: $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}^T$.
- This solution is optimal given the constraint $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}^T$.

Generalizing approximations

Nyström:

Expand the upper part:

$$\widetilde{\mathbf{U}}_{\mathbf{M}} = \begin{pmatrix} \mathbf{U}_{\mathbf{A}} \\ \mathbf{B}_{21}\mathbf{U}_{\mathbf{A}}\mathbf{\Lambda}_{\mathbf{A}}^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{U}_{\mathbf{A}}\mathbf{\Lambda}_{\mathbf{A}}^{-1} \\ \mathbf{B}_{21}\mathbf{U}_{\mathbf{A}}\mathbf{\Lambda}_{\mathbf{A}}^{-1} \end{pmatrix} = \overbrace{\mathbf{C}}^{N \times L}\underbrace{\mathbf{U}_{\mathbf{A}}\mathbf{\Lambda}_{\mathbf{A}}^{-1}}_{L \times d}$$

Column Sampling:

Rewrite using the eigendecomposition of $L \times L$ matrix $\mathbf{C}^T\mathbf{C}$:

$$\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{U}_{\mathbf{C}} = \mathbf{C}\mathbf{V}_{\mathbf{C}}\mathbf{\Sigma}_{\mathbf{C}}^{-1} = \mathbf{C}\mathbf{U}_{\mathbf{C}^T\mathbf{C}}\mathbf{\Lambda}_{\mathbf{C}^T\mathbf{C}}^{-1/2}$$

LLL:

Rewrite the solution $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}^T$ as $\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{Z}\widetilde{\mathbf{X}}^T$, where $\widetilde{\mathbf{X}}$ is computed optimally (given $\mathbf{Z}$) as:

$$\widetilde{\mathbf{X}} = \mathrm{eig}(\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}\mathbf{Z}^T)$$

Generalizing approximations

Nyström:

1. Solve the smaller $L \times L$ eigendecomposition:

$$\mathbf{A} = \mathbf{U_A} \mathbf{\Lambda_A} \mathbf{U_A}^T$$

2. Apply $N \times L$ out-of-sample matrix:

$$\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{C} \mathbf{U_A} \mathbf{\Lambda_A}^{-1}$$

Column Sampling:

1. Solve the smaller $L \times L$ eigendecomposition:

$$\mathbf{C}^T\mathbf{C} = \mathbf{U}_{\mathbf{C}^T\mathbf{C}} \mathbf{\Lambda}_{\mathbf{C}^T\mathbf{C}} \mathbf{U}_{\mathbf{C}^T\mathbf{C}}$$

2. Apply $N \times L$ out-of-sample matrix:

$$\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{C} \mathbf{U}_{\mathbf{C}^T\mathbf{C}} \mathbf{\Lambda}_{\mathbf{C}^T\mathbf{C}}^{-1/2}$$

LLL:

1. Solve the smaller $L \times L$ eigendecomposition:

$$\widetilde{\mathbf{X}} = \mathrm{eig}(\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}\mathbf{Z}^T)$$

2. Apply $N \times L$ out-of-sample matrix:

$$\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{Z}\widetilde{\mathbf{X}}^T$$

Generalizing approximations

Each approximation consist of the following steps:

- define an out-of-sample matrix $\mathbf{Z}_{N \times L}$,
- compute some reduced eigenproblem and a matrix $\mathbf{Q}_{L \times d}$ that depends on it,
- final approximation is equal to $\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{Z}\mathbf{Q}$.

	$\mathbf{Z}_{N \times L}$	Eigenproblem $\mathcal{A}\mathbf{U} = \mathcal{B}\mathbf{U}\boldsymbol{\Lambda}$ $\mathcal{A}, \mathcal{B}$	$\mathbf{Q}_{L \times d}$
Nyström	$\mathbf{C}$	$\mathbf{A}, \mathbf{I}$	$\mathbf{U}\boldsymbol{\Lambda}^{-1}$
Column Sampling	$\mathbf{C}$	$\mathbf{Z}^T\mathbf{Z}, \mathbf{I}$	$\mathbf{U}\boldsymbol{\Lambda}^{-1/2}$
LLL	computed from $\mathbf{Y} \approx \widetilde{\mathbf{Y}}\mathbf{Z}$	$\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}^T\mathbf{Z}$	$\mathbf{U}$
Random Projection	$\mathrm{qr}(\mathbf{M}^q\mathbf{S})$	$\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}^T\mathbf{Z}$	$\mathbf{U}$

Generalizing approximations

Each approximation consist of the following steps:

- define an out-of-sample matrix $\mathbf{Z}_{N \times L}$,
- compute some reduced eigenproblem and a matrix $\mathbf{Q}_{L \times d}$ that depends on it,
- final approximation is equal to $\widetilde{\mathbf{U}}_{\mathbf{M}} = \mathbf{Z}\mathbf{Q}$.

	$\mathbf{Z}_{N \times L}$	Eigenproblem $\mathcal{A}\mathbf{U} = \mathcal{B}\mathbf{U}\mathbf{\Lambda}$ $\mathcal{A}, \mathcal{B}$	$\mathbf{Q}_{L \times d}$
Nyström	$\mathbf{C}$	$\mathbf{A}, \mathbf{I}$	$\mathbf{U}\mathbf{\Lambda}^{-1}$
Column Sampling	$\mathbf{C}$	$\mathbf{Z}^T\mathbf{Z}, \mathbf{I}$	$\mathbf{U}\mathbf{\Lambda}^{-1/2}$
LLL	computed from $\mathbf{Y} \approx \widetilde{\mathbf{Y}}\mathbf{Z}$	$\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}^T\mathbf{Z}$	$\mathbf{U}$
Random Projection	$\mathrm{qr}(\mathbf{M}^q\mathbf{S})$	$\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}^T\mathbf{Z}$	$\mathbf{U}$
Variational Nyström	$\mathbf{C}$	$\mathbf{Z}\mathbf{M}\mathbf{Z}^T, \mathbf{Z}\mathbf{Z}^T$	$\mathbf{U}$

Variational Nyström

Add this Nyström out-of-sample constraint to the spectral problem:

$$\min_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}\mathbf{M}\mathbf{X}^T\right) \quad \text{s.t.} \quad \mathbf{X}\mathbf{X}^T = \mathbf{I},\ \mathbf{X} = \widetilde{\mathbf{X}}\mathbf{C}^T$$

$$\Downarrow$$

$$\min_{\widetilde{\mathbf{X}}} \operatorname{tr}\left(\widetilde{\mathbf{X}}\mathbf{C}^T\mathbf{M}\mathbf{C}\widetilde{\mathbf{X}}^T\right) \quad \text{s.t.} \quad \widetilde{\mathbf{X}}\mathbf{C}^T\mathbf{C}\widetilde{\mathbf{X}}^T = \mathbf{I}$$

From LLL perspective:

- replace customary built out-of-sample matrix $\mathbf{Z}$ with a readily available column matrix $\mathbf{C}$,
- abandon local linearity assumption of the weights $\mathbf{Z}$,
- save computation of $\mathbf{Z}$,
- $\mathbf{Z}$ is usually sparser than $\mathbf{C}$ (due to locality).

Variational Nyström

Add this Nyström out-of-sample constraint to the spectral problem:

$$\min_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}\mathbf{M}\mathbf{X}^T\right) \quad \text{s.t.} \quad \mathbf{X}\mathbf{X}^T = \mathbf{I},\ \mathbf{X} = \widetilde{\mathbf{X}}\mathbf{C}^T$$

$$\Downarrow$$

$$\min_{\widetilde{\mathbf{X}}} \operatorname{tr}\left(\widetilde{\mathbf{X}}\mathbf{C}^T\mathbf{M}\mathbf{C}\widetilde{\mathbf{X}}^T\right) \quad \text{s.t.} \quad \widetilde{\mathbf{X}}\mathbf{C}^T\mathbf{C}\widetilde{\mathbf{X}}^T = \mathbf{I}$$

From Nyström perspective:

- use the same out-of-sample matrix $\mathbf{C}$, but optimize the choice of the reduced eigenproblem,
- for fixed $\widetilde{\mathbf{Y}}$ gives better approx. than Nyström or Column Sampling (*optimal* for the out-of-sample kernel $\mathbf{C}$).
- uses all the elements from $\mathbf{M}$ to construct the reduced eigenproblem,
- forgo the interpolating property of Nyström.

Subsampling graph Laplacian

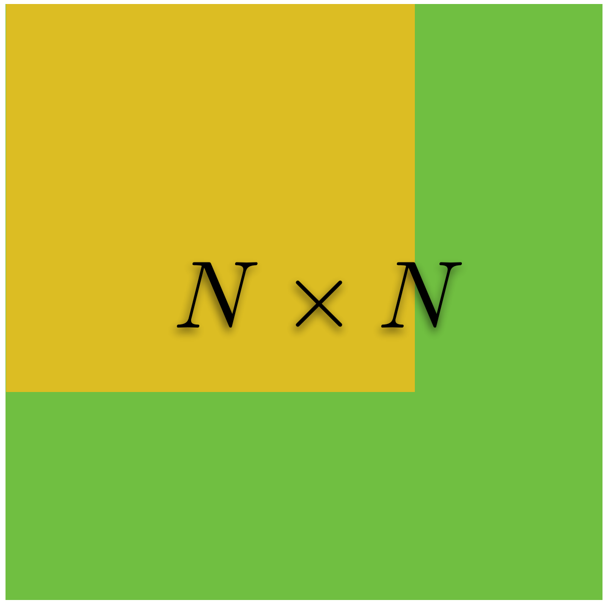
- Consider $\mathbf{M}$ given by normalized graph Laplacian matrix:

$$\mathbf{L} \propto \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$$

  - Gaussian affinity matrix: $w_{nm} = \exp(-\|\mathbf{y}_n^2 - \mathbf{y}_m^2\|/2\sigma^2)$
  - Degree matrix: $\mathbf{D} = \mathrm{diag}\left(\sum_{m=1}^{N} w_{nm}\right)$

- One of the most widely used kernel (Laplacian Eigenmaps, spectral clustering).
- Graph Laplacian kernel is a *data dependent*:

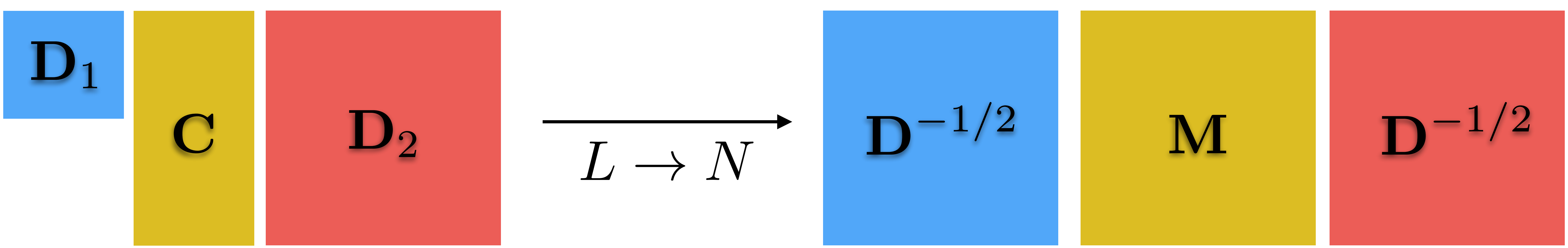
graph Laplacian computed for a subset of L input points $\neq$ $L \times L$ subset of graph Laplacian constructed for N points.



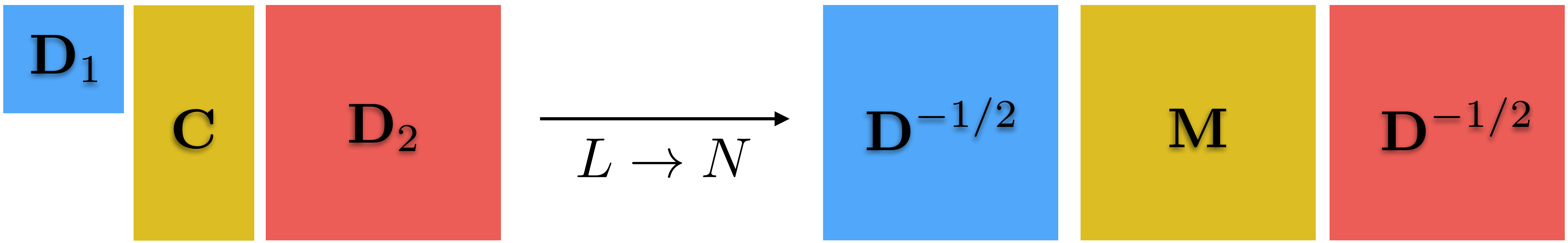
Subsampling graph Laplacian

- Data dependance can be a problem for methods that depend on the subsampling:
  - Nyström,
  - Column Sampling,
  - Variational Nyström.
- Not a problem methods for which there is no subsampling:
  - LLL,
  - Random projection.

Our solution: normalize subsample kernel separately, but in a way that interpolates over the landmarks and gives exact solution when $L = N$:



Subsampling graph Laplacian

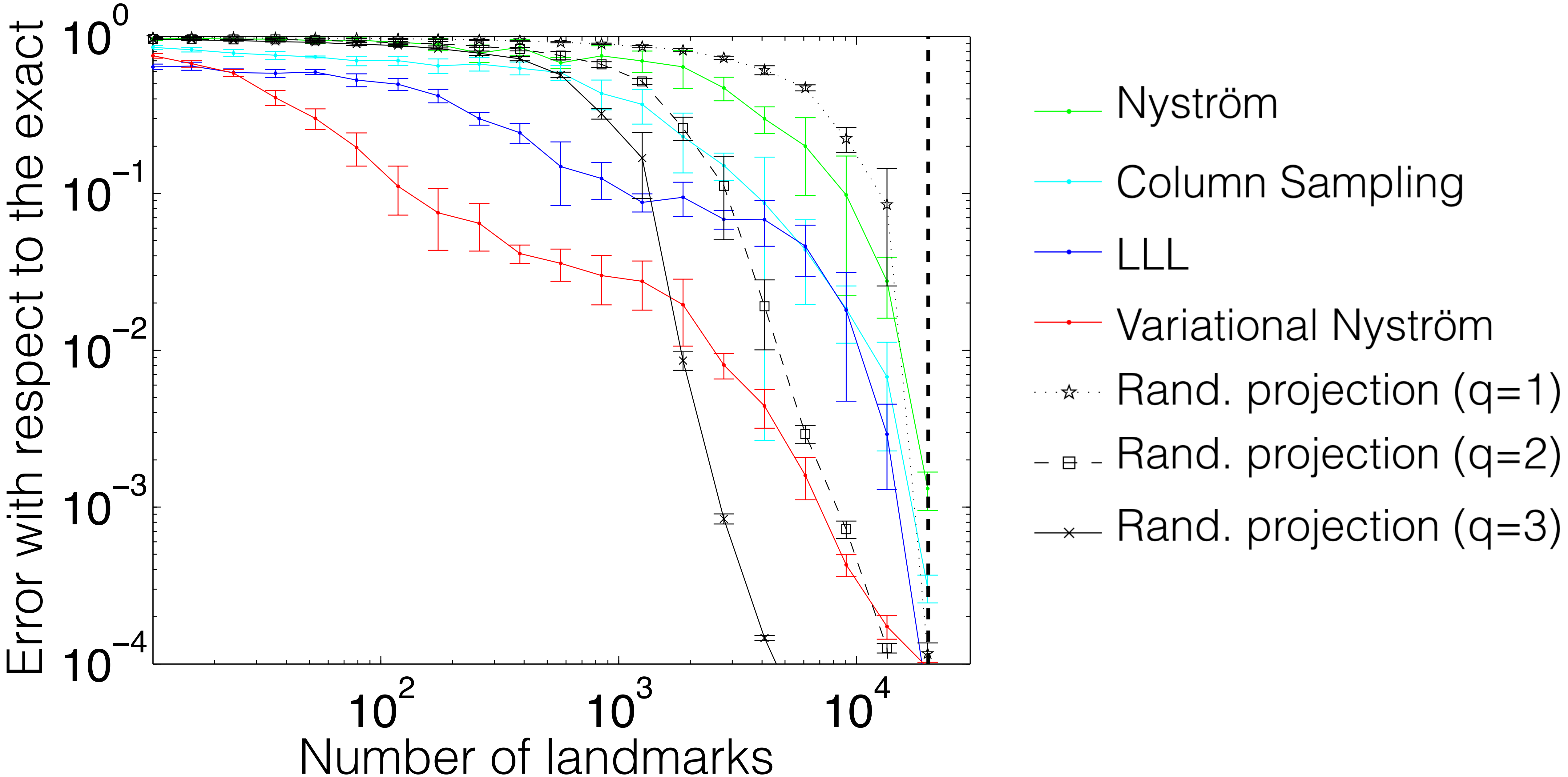


- For Nyström and Column Sampling:
  - we propose different forms for $\mathbf{D}_1$ and $\mathbf{D}_2$,
  - we evaluate empirically which one is the best.
- For Variational Nyström:
  - we showed that $\mathbf{D}_2$ factors out,
  - any $\mathbf{D}_1$ leads to the exact solution when $L = N$.

For the graph Laplacian kernel, the Variational Nyström approximation is more general.

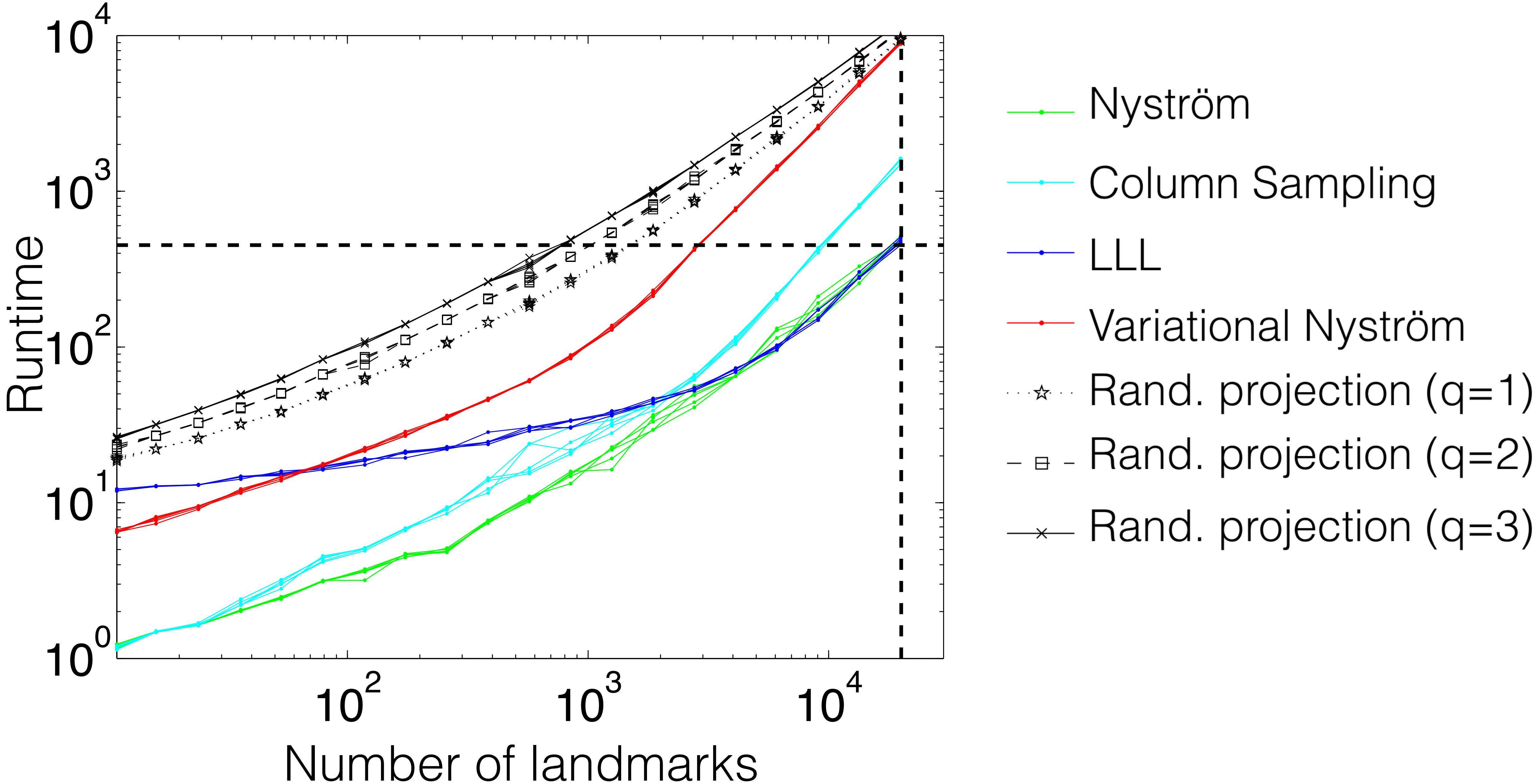
Experiments: Laplacian eigenmaps

- Reduce dimensionality of $N = 20\,000$ digits from MNIST $d = 10$.
- Run 5 times for different randomly chosen landmarks from $L = 11$ to $L = 19\,900$.



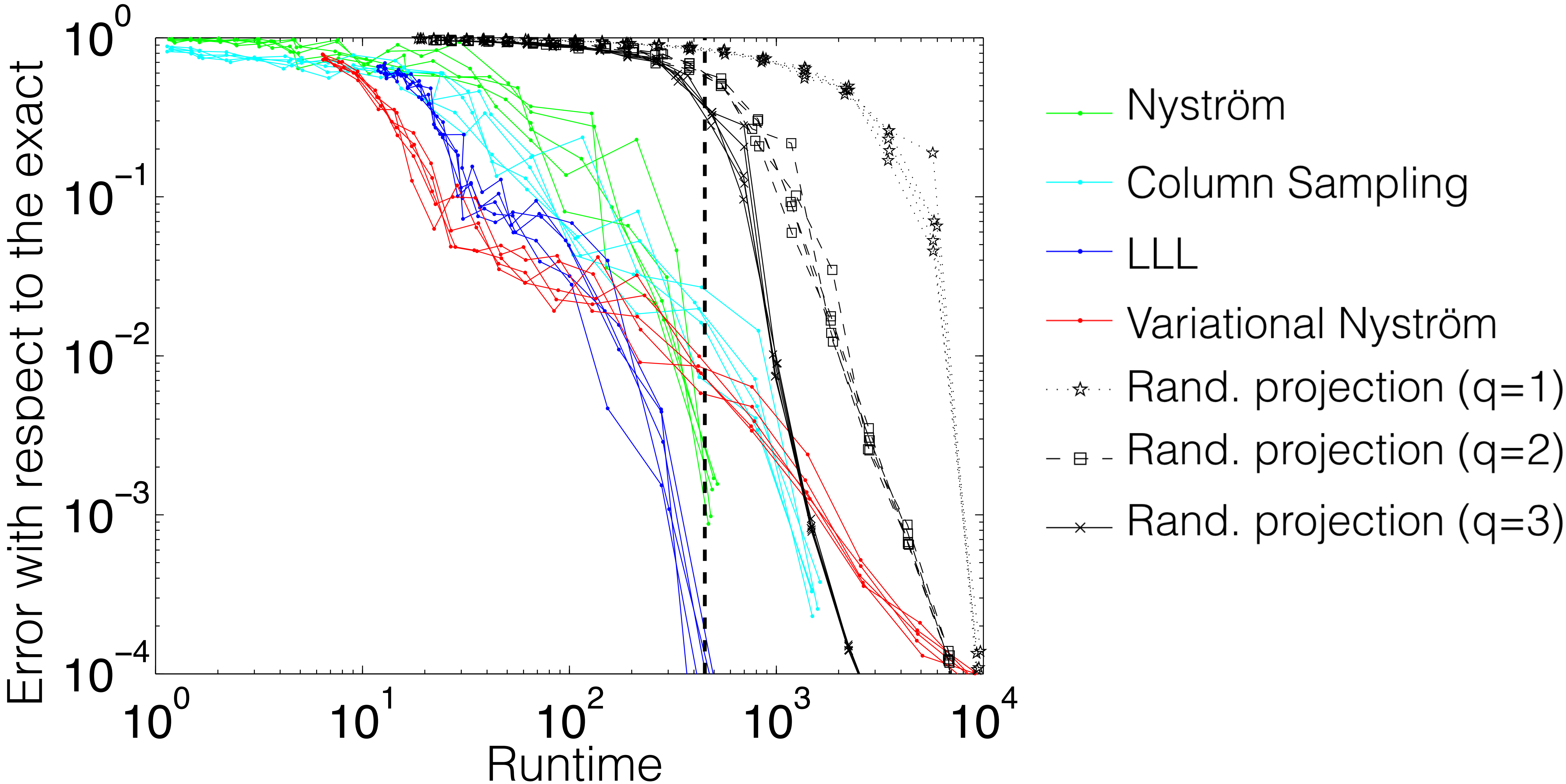
Experiments: Laplacian eigenmaps

- Reduce dimensionality of $N = 20\,000$ digits from MNIST $d = 10$.
- Run 5 times for different randomly chosen landmarks from $L = 11$ to $L = 19\,900$.



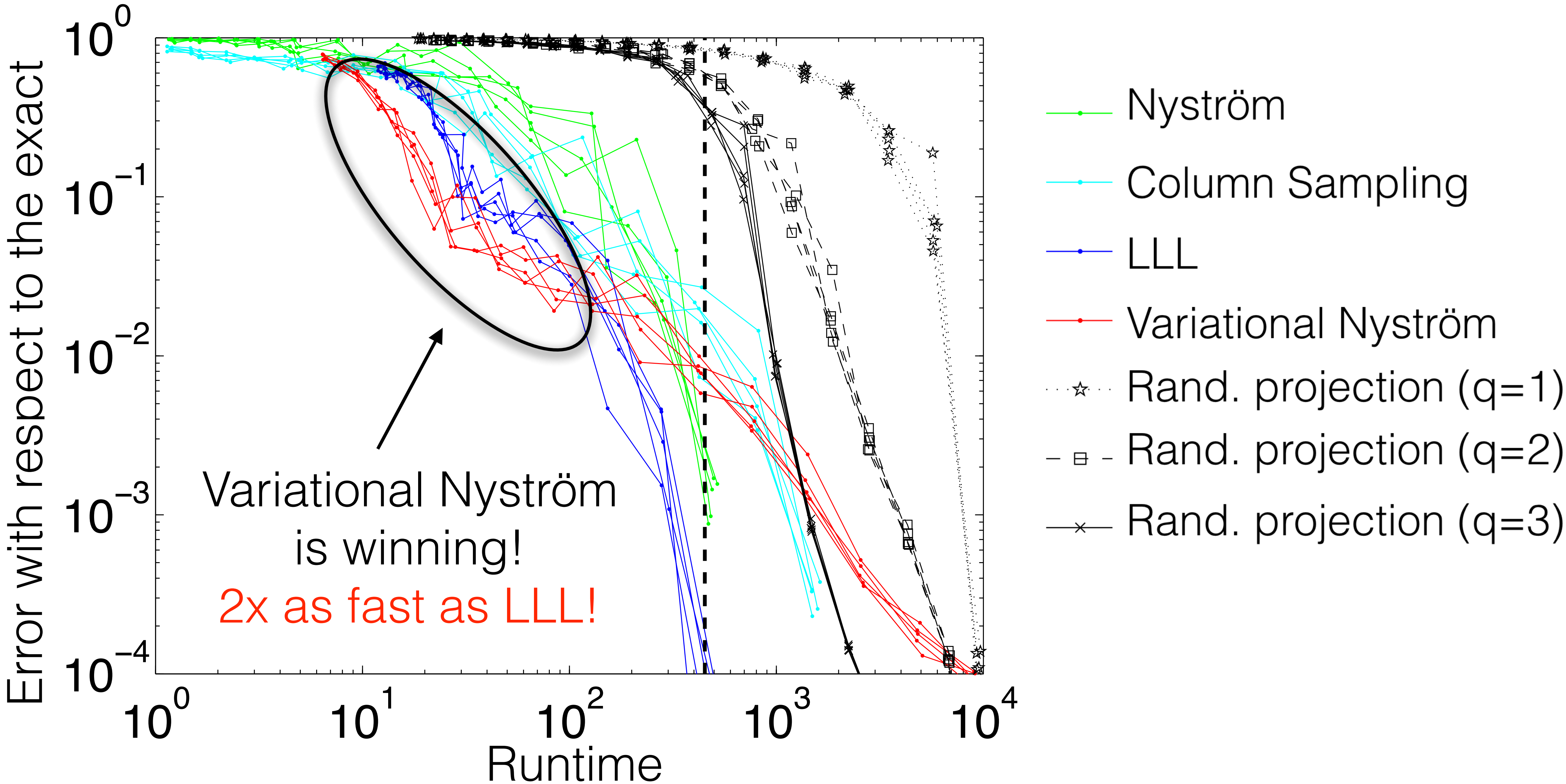
Experiments: Laplacian eigenmaps

- Reduce dimensionality of $N = 20\,000$ digits from MNIST $d = 10$.
- Run 5 times for different randomly chosen landmarks from $L = 11$ to $L = 19\,900$.



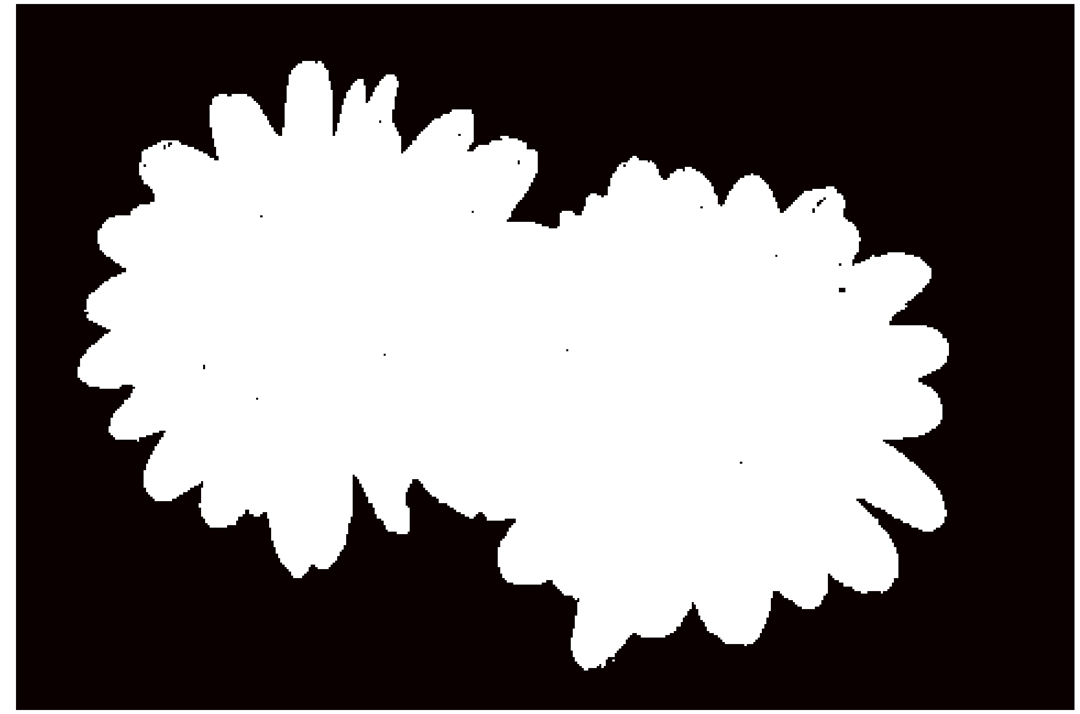
Experiments: Laplacian eigenmaps

- Reduce dimensionality of $N = 20\,000$ digits from MNIST $d = 10$.
- Run 5 times for different randomly chosen landmarks from $L = 11$ to $L = 19\,900$.

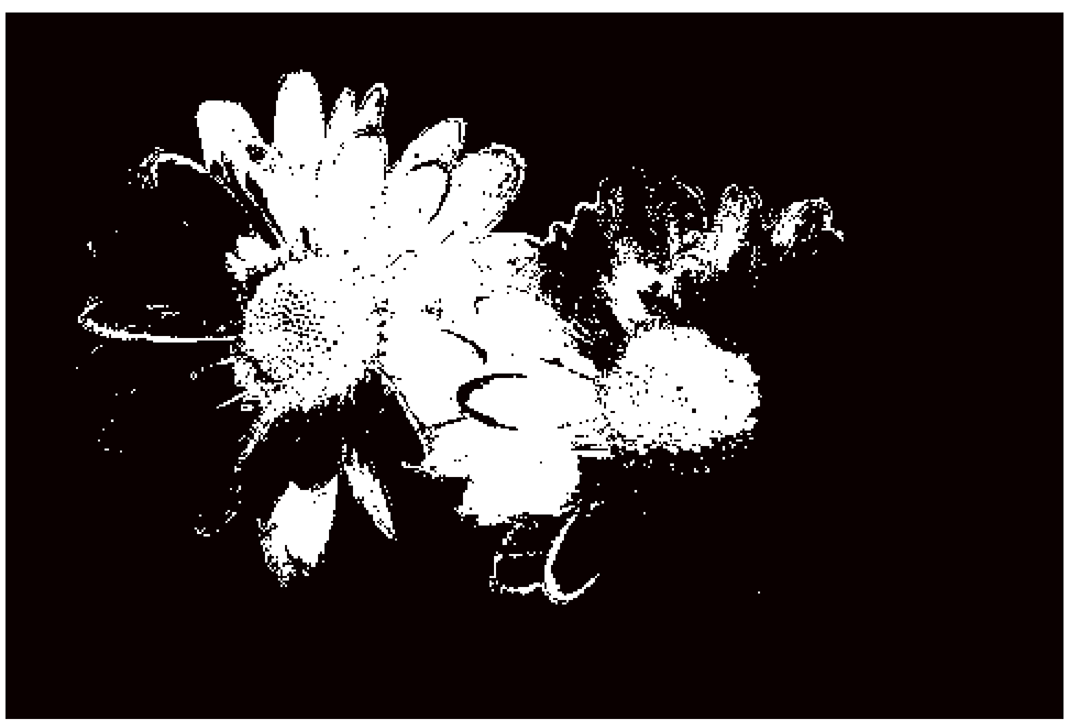


Experiments: Spectral clustering

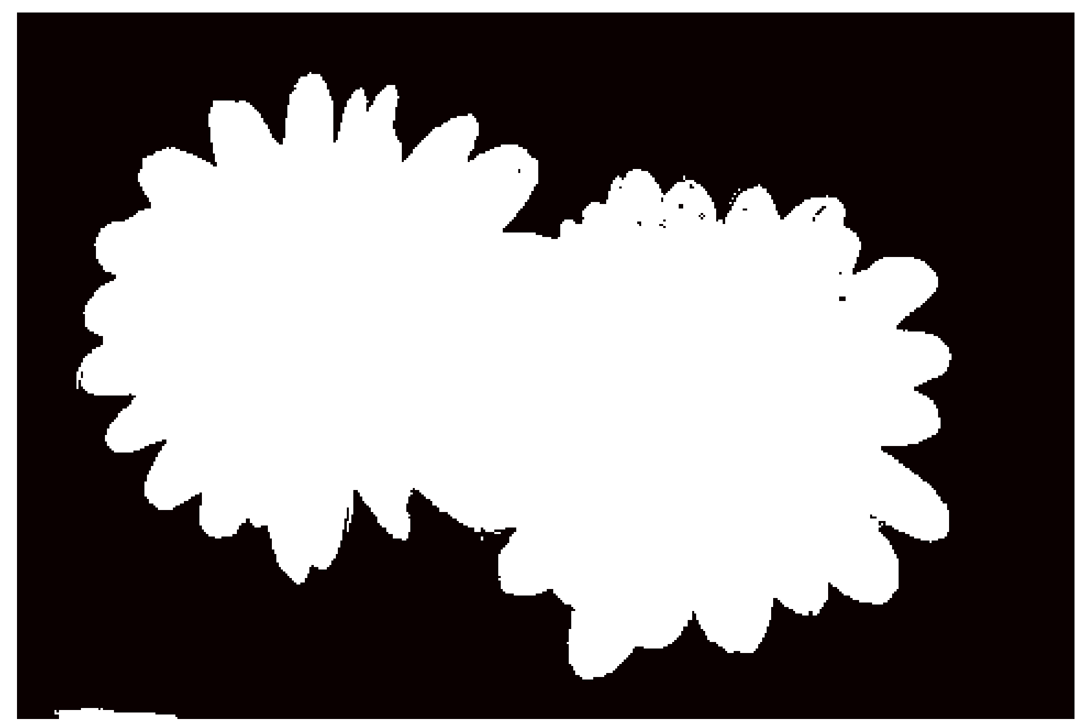
Original image



Exact Spectral clustering, $t = 512s$



Nyström, $t = 25s$

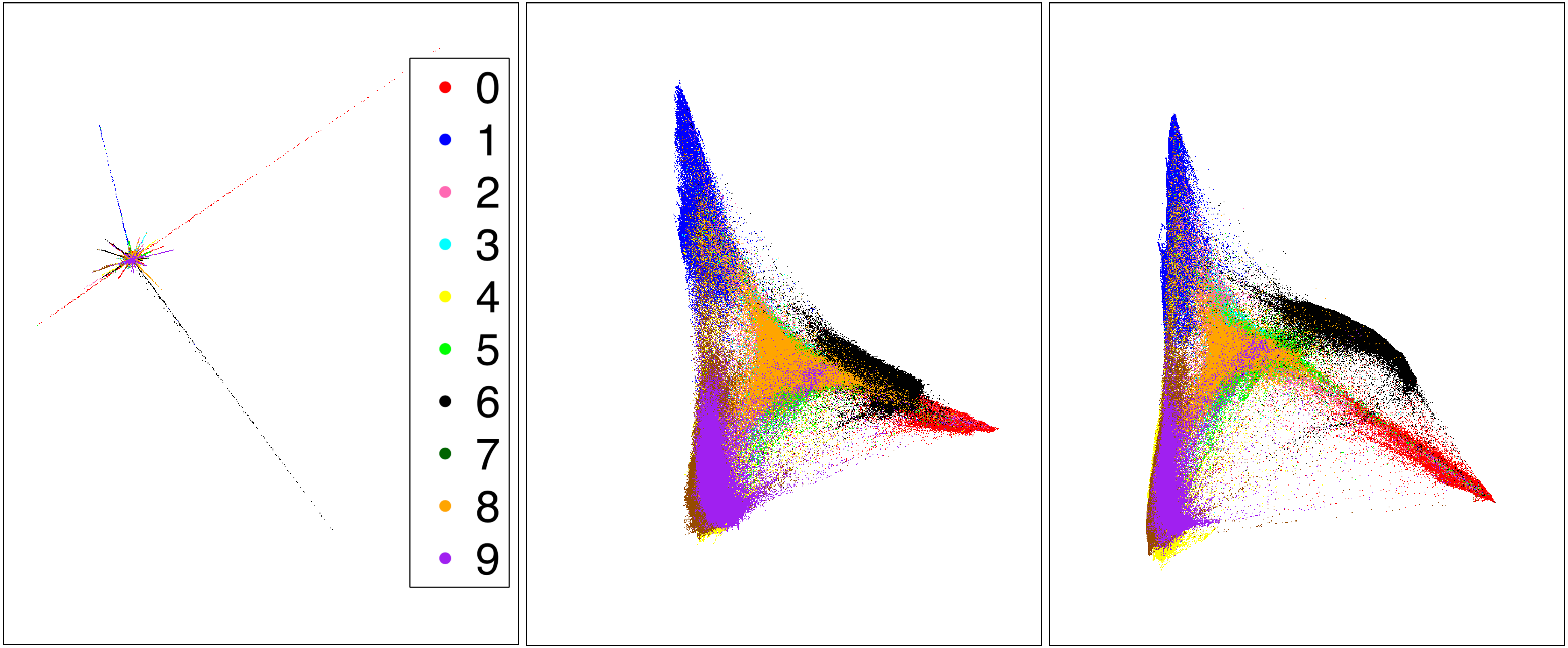


Variational Nyström, $t = 25s$

20x speedup!

infiniteMNIST embedding

Embedding of $N = 1\,020\,000$ digits from MNIST. Fix the runtime to $t = 10$ min



Nyström $L = 16\,000$

LLL $L = 5\,000$

Variational Nyström $L = 4\,500$

Conclusions

- The Variational Nyström method is the optimal way to use the out-of-sample Nyström formula to solve an eigenproblem approximately. It is able to achieve a low-to-medium accuracy solution faster than Nyström and other methods.
- We present a simple unified model of spectral clustering approximations, combining many existing algorithms such as Nyström, Column Sampling, LLL.
- We study the role of normalization in subsampling of the graph Laplacian kernel and show that Variational Nyström is more general for this kernel.

Poster #64 tomorrow (10am-1pm)

Partially supported by NSF award IIS-1423515