THE VARIATIONAL NYSTRÖM METHOD FOR LARGE-SCALE SPECTRAL PROBLEMS

Max Vladymyrov, Google Inc, and **Miguel Á. Carreira-Perpiñán,** EECS, UC Merced

1 Abstract

Spectral methods for dimensionality reduction and clustering require solving an eigenproblem defined by a sparse affinity matrix. When this matrix is large, one seeks an approximate solution. The standard way to do this is the Nyström method, which first solves a small eigenproblem considering only a subset of landmark points, and then applies an out-of-sample formula to extrapolate the solution to the entire dataset. We show that by constraining the original problem to satisfy the Nyström formula, we obtain an approximation that is computationally simple and efficient, but achieves a lower approximation error using fewer landmarks and less runtime. We also study the role of normalization in the computational cost and quality of the resulting solution.

2 Spectral methods

Consider a spectral problem:

$$\min_{\mathbf{X}} \operatorname{tr}\left(\mathbf{X}\mathbf{M}\mathbf{X}^T\right) \quad \text{s.t.} \quad \mathbf{X}\mathbf{X}^T = \mathbf{I} \tag{1}$$

where

- $\mathbf{M}$ is an $N \times N$ symmetric matrix (usually, a graph Laplacian) constructed on a dataset $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ of $D \times N$,
- $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ are coordinates in $\mathbb{R}^d$ for the N data points (embedding), $d < D$.

This minimization problem occurs in Laplacian Eigenmaps, ISOMAP, Locally Linear Embedding, Spectral Clustering, Kernel PCA etc.
The solution is given by the d trailing eigenvectors $\mathbf{U}_\mathbf{M}$ of $\mathbf{M}$, which is costly to compute when N is large. Our goal is to solve problems of the type (1) approximately.

3 Landmark approximation methods

Select L landmarks $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{y}}_1, \dots, \widetilde{\mathbf{y}}_L)$ from $\mathbf{Y}$ (e.g. randomly).
W.l.o.g. rearrange $\mathbf{M}$ with landmarks first: $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$ and $\mathbf{C} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B}_{21} \end{pmatrix}$ are the columns of $\mathbf{M}$ that correspond to $\widetilde{\mathbf{Y}}$.
Ways to approximate the computation of (1):

1. Nyström method

Essentially, an out-of-sample formula:

1. Solve the eigenproblem for $\widetilde{\mathbf{Y}}$.
2. Predict the rest of the points with an *interpolation* formula.

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

Problem: ignores non-landmark points for the initial prediction → interpolation over the bad solution is bad.

2. Column Sampling

Approximation is given by the left singular vectors of $\mathbf{C}$:
$\mathbf{C} = \mathbf{U}_\mathbf{C}\mathbf{\Sigma}_\mathbf{C}\mathbf{V}_\mathbf{C}^T \quad \Rightarrow \quad \widetilde{\mathbf{U}}_\mathbf{M} = \mathbf{U}_\mathbf{C}$
Rewriting using the SVD of $\mathbf{C}^T\mathbf{C}$ we get $\widetilde{\mathbf{U}}_\mathbf{M} = \mathbf{C}\mathbf{U}_{\mathbf{C}^T\mathbf{C}}\mathbf{\Lambda}_{\mathbf{C}^T\mathbf{C}}^{-1/2}$.
Problem: uses more data than Nyström (columns of $\mathbf{M}$), but still ignores non-landmark/non-landmark interactions $\mathbf{B}_{22}$ for prediction.

3. Locally Linear Landmarks

1. Construct the local linear projection matrix $\mathbf{Z}$ from $\mathbf{Y}$:
$$\mathbf{y}_n \approx \sum_{l=1}^{L} z_{ln}\widetilde{\mathbf{y}}_l, n = 1, \dots, N \quad \Rightarrow \quad \mathbf{Y} \approx \widetilde{\mathbf{Y}}\mathbf{Z} \tag{2}$$
2. Assume that projection is also satisfied in the embedding space: $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}$
3. Adding this constraint to (1) results in a reduced $L \times L$ eigenproblem:
$$\min_{\widetilde{\mathbf{X}}} \operatorname{tr}\left(\widetilde{\mathbf{X}}\widetilde{\mathbf{M}}\widetilde{\mathbf{X}}^T\right) \quad \text{s.t.} \quad \widetilde{\mathbf{X}}\mathbf{Z}\mathbf{Z}^T\widetilde{\mathbf{X}}^T = \mathbf{I}$$
with reduced affinities $\widetilde{\mathbf{M}} = \mathbf{Z}\mathbf{M}\mathbf{Z}^T$.
4. Reconstruct final embedding using $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{Z}$

Final eigenvector approximation: $\widetilde{\mathbf{U}}_\mathbf{M} = \mathbf{Z}\widetilde{\mathbf{X}}$.
Problem: Local linearly assumption may not be always true. Strong dependance on features $\mathbf{Y}$.

4. Random Projection method

1. Use a random matrix $\mathbf{S}_{N \times L}$ to form a low-dimensional sample matrix $\mathbf{M}_\mathbf{S} = \mathbf{M}\mathbf{S}$,
2. Compute SVD of the projection of $\mathbf{M}$ onto QR decomposition of $\mathbf{M}_\mathbf{S}$,
3. Project the results back to the original space.

4 Generalizing approximations

All the methods above can be generalized as follows:

1. Define an out-of-sample matrix $\mathbf{Z}$.
2. Compute a reduced eigenproblem and a matrix $\mathbf{Q}_{L \times d}$ that depends on it.
3. Final approximation is equal to $\mathbf{U}_\mathbf{M} = \mathbf{Z}\mathbf{Q}$

Algorithm	$\mathbf{Z}_{N \times L}$	$\mathbf{Q}_{L \times d}$	Eigenproblem $\mathcal{A}\mathbf{U} = \mathcal{B}\mathbf{U}\mathbf{\Lambda}$ $\mathcal{A}$, $\mathcal{B}$
Nyström	$\mathbf{C}$	$\mathbf{U}\mathbf{\Lambda}^{-1}$	$\mathbf{A}$, $\mathbf{I}$
Column sampling	$\mathbf{C}$	$\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}$	$\mathbf{Z}^T\mathbf{Z}$, $\mathbf{I}$
Random Projection	$\operatorname{qr}(\mathbf{M}^q\mathbf{S})$	$\mathbf{U}$	$\mathbf{Z}^T\mathbf{M}\mathbf{Z}$, $\mathbf{I}$
LLL	eq. (2)	$\mathbf{U}$	$\mathbf{Z}^T\mathbf{M}\mathbf{Z}$, $\mathbf{Z}^T\mathbf{Z}$
Variational Nyström	$\mathbf{C}$	$\mathbf{U}$	$\mathbf{Z}^T\mathbf{M}\mathbf{Z}$, $\mathbf{Z}^T\mathbf{Z}$

5 Variational Nyström

Since the embedding is defined by the Nyström out-of-sample formula $\mathbf{X} = \widetilde{\mathbf{X}}\mathbf{C}^T$, we add this as a constraint to the spectral problem (1):

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

Which result in a reduced eigenproblem:

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

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 $\mathbf{Z}$ and dependance on the features $\mathbf{Y}$,
- save computation of $\mathbf{Z}$.

From Nyström perspective:

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

6 Subsampling Graph Laplacians

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

$$\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-1/2},$$

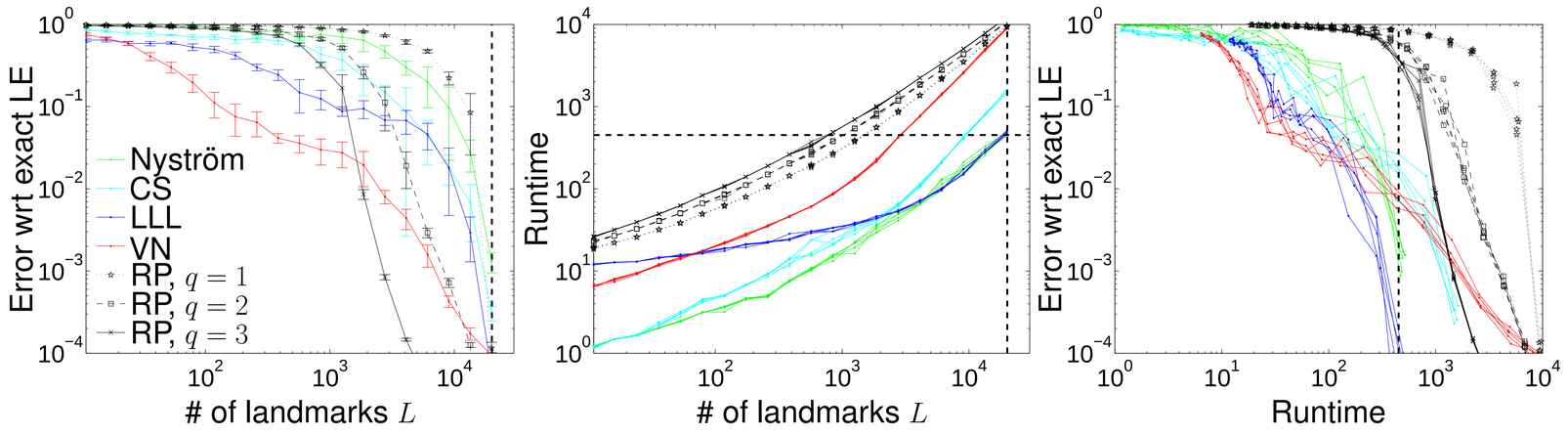
(as in Laplacian Eigenmaps, spectral clustring etc.) where

$$w_{nm} = \exp(-\|\mathbf{y}_n^2 - \mathbf{y}_m^2\|/2\sigma^2) \qquad \mathbf{D} = \operatorname{diag}\left(\sum_{m=1}^{N} w_{nm}\right)$$

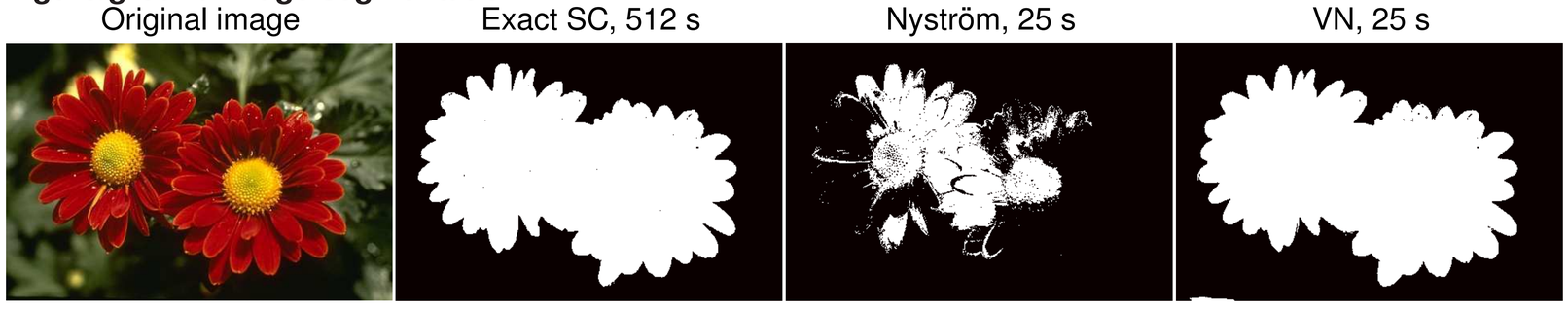
- $\mathbf{L}$ is a data dependent kernel: Graph Laplacian computed for a subset of L input points *is not equal* to the $L \times L$ subset of graph Laplacian constructed for N points.
- This can be *problematic* for methods that depend on subsampling, such as Nyström and Variational Nyström. Not a problem for LLL, since there is no subsampling involved.
- Our solution: normalize out-of-sample kernel separately, but in a way that (1) interpolates over the landmarks and (2) gives exact solution when $L = N$:
- We show that for the Variational Nyström final normalization is more general and has much simpler form than for Nyström method.

7 Experiments

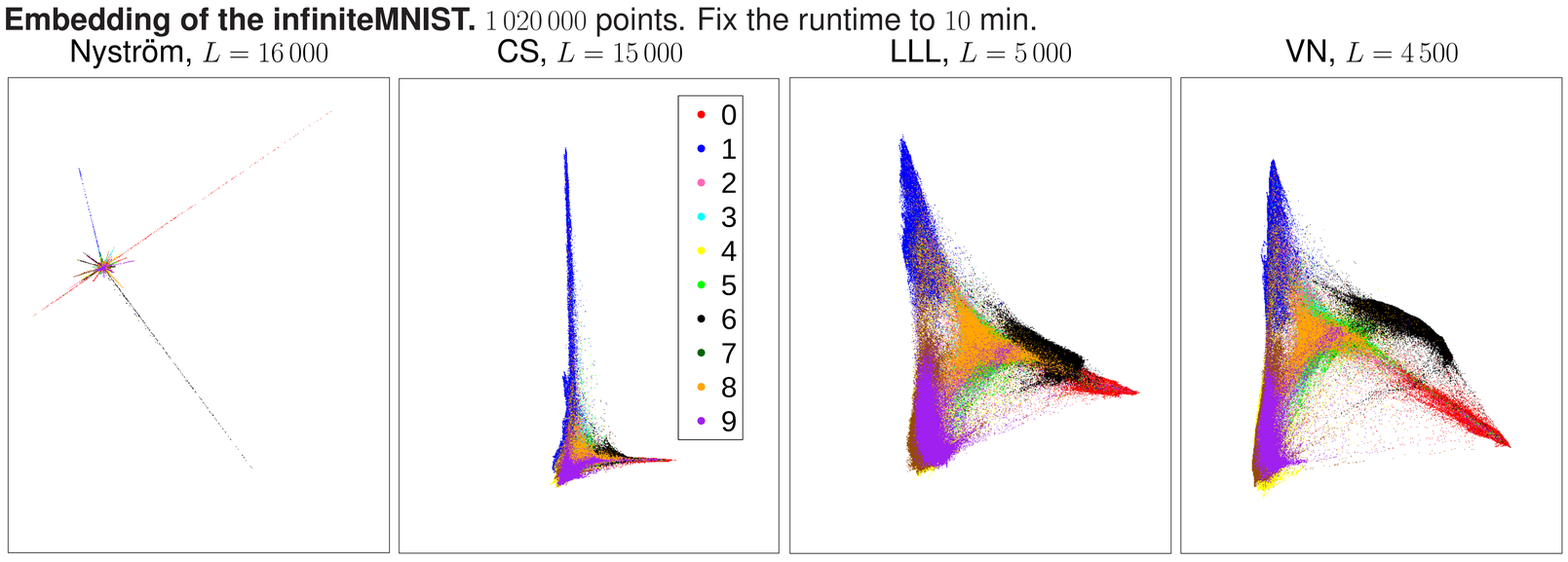
Medium size experiment. Reduce dimensionality of $20\,000$ random points from MNIST to $d = 10$.
Error/runtime tradeoff of different methods.



Figure/ground image segmentation.



Embedding of the infiniteMNIST. $1\,020\,000$ points. Fix the runtime to 10 min.



8 Conclusions

1. 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.
2. We present a simple unified model of spectral approximations, combining many existing algorithms such as Nyström, Column Sampling, LLL.
3. We study the role of normalization in subsampling of the graph Laplacian kernel.

ICML @ NYC
International Conference on Machine Learning

Partially supported by NSF award IIS–1423515.