A novel greedy algorithm for Nyström approximation Supplementary material

Ahmed K. Farahat **Ali Ghodsi** **Mohamed S. Kamel**
University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

1 UPDATE FORMULAS FOR $\boldsymbol{f}$ AND $\boldsymbol{g}$

In this section, the recursive formulas for $\boldsymbol{f}$ and $\boldsymbol{g}$ (Equation (18) in the main paper) are derived. The greedy selection criterion at iteration t is:

$$q = arg \max_i \| \frac{1}{\sqrt{E_{ii}}} E_{:i} \|^2, \tag{1}$$

where E is the residual matrix at iteration t, $E_{:i}$ denotes the i-th column of E, and E_{ii} denotes the i-th diagonal element of E.

As the Nyström approximation is calculated in a recursive manner based on the residual matrix at the previous iteration, E, $E_{:i}$, and E_{ii} for a candidate column i can be recursively calculated as follows:

$$\begin{aligned} E^{(t)} &= (E - \frac{1}{\alpha}\boldsymbol{\delta}\boldsymbol{\delta}^T)^{(t-1)} = (E - \boldsymbol{\omega}\boldsymbol{\omega}^T)^{(t-1)}, \\ E^{(t)}_{:i} &= (E_{:i} - \frac{\boldsymbol{\delta}_i}{\alpha}\boldsymbol{\delta})^{(t-1)} = (E_{:i} - \boldsymbol{\omega}_i\boldsymbol{\omega})^{(t-1)}, \\ E^{(t)}_{ii} &= (E_{ii} - \frac{\boldsymbol{\delta}_i^2}{\alpha})^{(t-1)} = (E_{ii} - \boldsymbol{\omega}_i^2)^{(t-1)}. \end{aligned} \tag{2}$$

Let $\boldsymbol{f}_i = \|E_{:i}\|^2$ and $\boldsymbol{g}_i = E_{ii}$ be the numerator and denominator of the criterion function for data point i respectively. Based on (2), $\boldsymbol{f}_i^{(t)}$ can be calculated as:

$$\begin{aligned} \boldsymbol{f}_i^{(t)} &= \left(\|E_{:i} - \boldsymbol{\omega}_i\boldsymbol{\omega}\|^2\right)^{(t-1)} \\ &= \left((E_{:i} - \boldsymbol{\omega}_i\boldsymbol{\omega})^T(E_{:i} - \boldsymbol{\omega}_i\boldsymbol{\omega})\right)^{(t-1)} \\ &= \left(E_{:i}^T E_{:i} - 2\boldsymbol{\omega}_i E_{:i}^T\boldsymbol{\omega} + \boldsymbol{\omega}_i^2\|\boldsymbol{\omega}\|^2\right)^{(t-1)} \\ &= \left(\boldsymbol{f}_i - 2\boldsymbol{\omega}_i E_{:i}^T\boldsymbol{\omega} + \boldsymbol{\omega}_i^2\|\boldsymbol{\omega}\|^2\right)^{(t-1)}. \end{aligned} \tag{3}$$

Similarly, $\boldsymbol{g}_i^{(t)}$ can be calculated as:

$$\begin{aligned} \boldsymbol{g}_i^{(t)} = E_{ii}^{(t)} &= \left(E_{ii} - \boldsymbol{\omega}_i^2\right)^{(t-1)} \\ &= \left(\boldsymbol{g}_i - \boldsymbol{\omega}_i^2\right)^{(t-1)}. \end{aligned} \tag{4}$$

Let $\boldsymbol{f} = [\boldsymbol{f}_i]_{i=1..n}$ and $\boldsymbol{g} = [\boldsymbol{g}_i]_{i=1..n}$, $\boldsymbol{f}^{(t)}$ and $\boldsymbol{g}^{(t)}$ can be expressed as:

$$\begin{aligned} \boldsymbol{f}^{(t)} &= \left(\boldsymbol{f} - 2\left(\boldsymbol{\omega} \circ E\boldsymbol{\omega}\right) + \|\boldsymbol{\omega}\|^2\left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\right)^{(t-1)}, \\ \boldsymbol{g}^{(t)} &= \left(\boldsymbol{g} - \left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\right)^{(t-1)}, \end{aligned} \tag{5}$$

where $\circ$ represents the Hadamard product operator, and $\|.\|$ is the ℓ_2 norm.

Based on the recursive formula of E, the term $E\boldsymbol{\omega}$ at iteration $(t-1)$ can be expressed as:

$$\begin{aligned} E\boldsymbol{\omega} &= \left(K - \Sigma_{r=1}^{t-2}\left(\boldsymbol{\omega}\boldsymbol{\omega}^T\right)^{(r)}\right)\boldsymbol{\omega} \\ &= K\boldsymbol{\omega} - \Sigma_{r=1}^{t-2}\left(\boldsymbol{\omega}^{(r)^T}\boldsymbol{\omega}\right)\boldsymbol{\omega}^{(r)} \end{aligned} \tag{6}$$

Substitute with $E\boldsymbol{\omega}$ in Equation (5), the update formulas for $\boldsymbol{f}$ and $\boldsymbol{g}$ are given as:

$$\begin{aligned} \boldsymbol{f}^{(t)} &= \Big(\boldsymbol{f} - 2\left(\boldsymbol{\omega} \circ \left(K\boldsymbol{\omega} - \Sigma_{r=1}^{t-2}\left(\boldsymbol{\omega}^{(r)^T}\boldsymbol{\omega}\right)\boldsymbol{\omega}^{(r)}\right)\right) \\ &\quad + \|\boldsymbol{\omega}\|^2\left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\Big)^{(t-1)}, \\ \boldsymbol{g}^{(t)} &= \Big(\boldsymbol{g} - \left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\Big)^{(t-1)}. \end{aligned} \tag{7}$$

In the case of partition-based greedy Nyström algorithm, the update formulas for $\boldsymbol{f}$ and $\boldsymbol{g}$ (Equation (19) in the main paper) can be derived as follows.

Let $E^{(t)}$ and $H^{(t)}$ be the residual matrices of K and G at iteration t respectively. The efficient sampling criterion based on centroids can be expressed as follows:

$$q = arg \max_i \| \frac{1}{\sqrt{E_{ii}}} H_{:i} \|^2, \tag{8}$$

where $H_{:i}$ denotes the i-th column of H, and E_{ii} denotes the i-th diagonal element of E. The term $H_{ji}/\sqrt{E_{ii}}$ is the scalar projection of the j-th centroid onto $X_{:i}$. Let $\boldsymbol{\delta}^{(t)}$ be column of E selected at iteration t, $\alpha^{(t)}$ be the corresponding diagonal element of E, and $\boldsymbol{\gamma}^{(t)}$ be the corresponding column of H. Define $\boldsymbol{\omega}^{(t)} = \boldsymbol{\delta}^{(t)}/\sqrt{\alpha^{(t)}}$, and $\boldsymbol{\upsilon}^{(t)} = \boldsymbol{\gamma}^{(t)}/\sqrt{\alpha^{(t)}}$. The rank-1 approximation of $H^{(t)}$ can be calculated as:

$$\tilde{H}^{(t)}_{\{q\}} = \frac{1}{\alpha^{(t)}}\boldsymbol{\gamma}^{(t)}\boldsymbol{\delta}^{T(t)} = \boldsymbol{\upsilon}^{(t)}\boldsymbol{\omega}^{T(t)}, \tag{9}$$

and the new residual matrix H can be calculated as:

$$H^{(t+1)} = H^{(t)} - \boldsymbol{\upsilon}^{(t)}\boldsymbol{\omega}^{T(t)} \tag{10}$$

Based on this recursive formula, the greedy sampling criterion (Equation 8) can be calculated in a recursive manner as follows. Similar to Equation (2), H, $H_{:i}$, and E_{ii} can be recursively calculated as:

$$\begin{aligned} H^{(t)} &= (H - \boldsymbol{v}\boldsymbol{\omega}^T)^{(t-1)}, \\ H_{:i}^{(t)} &= (H_{:i} - \boldsymbol{\omega}_i \boldsymbol{v})^{(t-1)}, \\ E_{ii}^{(t)} &= (E_{ii} - \boldsymbol{\omega}_i^2)^{(t-1)}. \end{aligned} \tag{11}$$

Let $\boldsymbol{f}_i = \|H_{:i}\|^2$ and $\boldsymbol{g}_i = E_{ii}$ be the numerator and denominator of the criterion function for data point i respectively. $\boldsymbol{f}_i^{(t)}$ and $\boldsymbol{g}_i^{(t)}$ can be calculated as follows:

$$\begin{aligned} \boldsymbol{f}_i^{(t)} &= \left(\|H_{:i} - \boldsymbol{\omega}_i \boldsymbol{v}\|^2\right)^{(t-1)} \\ &= \left(\boldsymbol{f}_i - 2\boldsymbol{\omega}_i H_{:i}^T \boldsymbol{v} + \boldsymbol{\omega}_i^2 \|\boldsymbol{v}\|^2\right)^{(t-1)}, \\ \boldsymbol{g}_i^{(t)} &= E_{ii}^{(t)} = \left(E_{ii} - \boldsymbol{\omega}_i^2\right)^{(t-1)} \\ &= \left(\boldsymbol{g}_i - \boldsymbol{\omega}_i^2\right)^{(t-1)}. \end{aligned} \tag{12}$$

Let $\boldsymbol{f} = [\boldsymbol{f}_i]_{i=1..n}$ and $\boldsymbol{g} = [\boldsymbol{g}_i]_{i=1..n}$, $\boldsymbol{f}^{(t)}$ and $\boldsymbol{g}^{(t)}$ can be expressed as:

$$\begin{aligned} \boldsymbol{f}^{(t)} &= \left(\boldsymbol{f} - 2\left(\boldsymbol{\omega} \circ H^T \boldsymbol{v}\right) + \|\boldsymbol{v}\|^2 \left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\right)^{(t-1)}, \\ \boldsymbol{g}^{(t)} &= \left(\boldsymbol{g} - \left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\right)^{(t-1)}, \end{aligned} \tag{13}$$

where $\circ$ represents the Hadamard product operator, and $\|.\|$ is the ℓ_2 norm.

The term $H^T\boldsymbol{v}$ at iteration $(t-1)$ can be calculated recursively as:

$$\begin{aligned} H^T \boldsymbol{v} &= \left(G^T - \Sigma_{r=1}^{t-2} \left(\boldsymbol{\omega}\boldsymbol{v}^T\right)^{(r)}\right) \boldsymbol{v} \\ &= G^T \boldsymbol{v} - \Sigma_{r=1}^{t-2} \left(\boldsymbol{v}^{(r)T} \boldsymbol{v}\right) \boldsymbol{\omega}^{(r)} \end{aligned} \tag{14}$$

Substitute with $H^T\boldsymbol{v}$ in Equation (13), the update formulas for $\boldsymbol{f}$ and $\boldsymbol{g}$ are given as:

$$\begin{aligned} \boldsymbol{f}^{(t)} &= \Big(\boldsymbol{f} - 2\Big(\boldsymbol{\omega} \circ \Big(G^T \boldsymbol{v} - \Sigma_{r=1}^{t-2} \left(\boldsymbol{v}^{(r)T} \boldsymbol{v}\right) \boldsymbol{\omega}^{(r)}\Big)\Big) \\ &\quad + \|\boldsymbol{v}\|^2 \left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\Big)^{(t-1)}, \\ \boldsymbol{g}^{(t)} &= \Big(\boldsymbol{g} - \left(\boldsymbol{\omega} \circ \boldsymbol{\omega}\right)\Big)^{(t-1)}. \end{aligned} \tag{15}$$

2 COMPARISON WITH ENSEMBLE NYSTRÖM

In this section, the proposed greedy Nyström methods (**GreedyNyström** and **PartGreedyNys**) are compared to the ensemble Nytröm algorithm (**EnsembleNyström**) proposed by Kumar et al. (2009). The ensemble Nyström method constructs a low-rank approximation of a kernel matrix using an ensemble of p Nyström approximations. As suggested by Kumar et al. (2009), the ridge regression algorithm can be used to learn the mixture weights of different approximations using a validation set of columns sampled from the original kernel matrix. In this experiment, an ensemble of $p = 10$ Nyström approximations is used, and l columns are sampled to calculate each low-rank approximation. A validation set of $s = 20$ columns is used for estimating the mixture weights of the ensemble, and a hold-out set of $s' = 20$ columns is used to estimate the ridge parameter.

Tables 1 and 2 show the relative accuracies and run time of different methods. Two values are used for l: $l = 3\%n$ and $l = 5\%n$, with $k = 1\%n$. It can be observed that both **PartGreedyNys** and **GreedyNyström** outperforms the ensemble method (**EnsembleNyström**) in term of approximation accuracy for most data sets.

References

S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström Method. In *Advances in Neural Information Processing Systems 22*, pages 1060–1068, 2009.

Table 1: The relative accuracy of rank-k approximations $\tilde{K}_{\mathcal{S},k}$ ($k = 1\%n$) for the proposed greedy Nyström methods compared to the ensemble method and Nyström method with uniform sampling. The best method for each data set is highlighted in bold, the second best method is underlined.

Datasets	UniNoRep	EnsembleNyström	GreedyNyström	PartGreedyNys-c=100
l/n= **3%**				
Reuters-21578	0.5306	0.5965	**0.9074**	<u>0.8897</u>
Reviews	0.7329	0.8383	<u>0.9096</u>	**0.9135**
LA1	0.8194	**0.9320**	<u>0.9302</u>	0.9294
MNIST-4K	0.5749	0.7199	**0.8476**	<u>0.8174</u>
PIE-20	0.5382	0.6918	**0.9102**	<u>0.8848</u>
Yale-B-38	0.5676	0.7119	**0.8820**	<u>0.8655</u>
l/n= **5%**				
Reuters-21578	0.6076	0.6593	**0.9603**	<u>0.9548</u>
Reviews	0.7998	0.8774	<u>0.9446</u>	**0.9454**
LA1	0.8546	0.9445	**0.9534**	<u>0.9523</u>
MNIST-4K	0.7131	0.8669	**0.9433**	<u>0.9314</u>
PIE-20	0.6797	0.8492	**0.9792**	<u>0.9735</u>
Yale-B-38	0.6937	0.8435	**0.9637**	<u>0.9562</u>

Table 2: The run-times (in seconds) corresponding to the relative accuracies shown in Table 1.

Datasets	UniNoRep	EnsembleNyström	GreedyNyström	PartGreedyNys-c=100
l/n= **3%**				
Reuters-21578	1.63	33.29	17.24	4.82
Reviews	0.86	26.02	12.61	2.31
LA1	0.57	18.54	6.41	1.21
MNIST-4K	0.79	23.93	11.98	2.13
PIE-20	0.59	19.72	7.58	1.40
Yale-B-38	0.29	11.64	2.89	0.65
l/n= **5%**				
Reuters-21578	1.73	35.86	28.53	6.79
Reviews	0.87	27.04	21.38	4.08
LA1	0.63	19.52	10.84	1.95
MNIST-4K	0.88	24.83	21.19	3.87
PIE-20	0.57	20.28	12.77	2.35
Yale-B-38	0.28	11.98	4.69	0.82