Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention (Supplementary Material)

Yunyang Xiong [1] **Zhanpeng Zeng** [1] **Rudrasis Chakraborty** [2] **Mingxing Tan** [3]
Glenn Fung [4] **Yin Li** [1] **Vikas Singh** [1]
[1] University of Wisconsin-Madison [2] UC Berkeley [3] Google Brain [4] American Family Insurance
yxiong43@wisc.edu, zzeng38@wisc.edu, rudra@berkeley.edu, tanmingxing@google.com, gfung@amfam.com,
yin.li@wisc.edu, vsingh@biostat.wisc.edu

This supplementary material derives standard Nyström approximation form for softmax matrix in detail, presents our proofs of theorems, describes our implementation details, and provides further experimental results.

Nyström Approximation for Softmax Matrix

Denote the softmax matrix used in self-attention $S = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right) \in \mathbf{R}^{n\times n}$. S can be written as

$$S = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right) = \begin{bmatrix} A_S & B_S \\ F_S & C_S \end{bmatrix}, \tag{1}$$

where $A_S \in \mathbf{R}^{m\times m}, B_S \in \mathbf{R}^{m\times(n-m)}, F_S \in \mathbf{R}^{(n-m)\times m}$ and $C_S \in \mathbf{R}^{(n-m)\times(n-m)}$. A_S is designated to be our sample matrix by sampling m columns and rows from S. The singular value decomposition (SVD) of the sample matrix can be written as, $A_S = U\Lambda V^T$, where $U, V \in \mathbf{R}^{m\times m}$ are orthogonal matrices, $\Lambda \in \mathbf{R}^{m\times m}$ is a diagonal matrix.

Given a query q_i and key k_j, let

$$\mathcal{K}_K(q_i) = \text{softmax}\left(\frac{q_i K^T}{\sqrt{d_q}}\right);\ \mathcal{K}_Q(k_j) = \text{softmax}\left(\frac{Qk_j^T}{\sqrt{d_q}}\right)$$

where $\mathcal{K}_K(q_i) \in \mathbf{R}^{1\times n}$ and $\mathcal{K}_Q(k_j) \in \mathbf{R}^{n\times 1}$. We can then construct

$$\phi_K(q_i) = \Lambda^{-\frac{1}{2}} V^T [\mathcal{K}_K^T(q_i)]_{m\times 1}$$
$$\phi_Q(k_j) = \Lambda^{-\frac{1}{2}} U^T [\mathcal{K}_Q(k_j)]_{m\times 1}$$

where $[\cdot]_{m\times 1}$ refers to calculating the full $n\times 1$ vector and then taking the first $m\times 1$ entries. With $\phi_K(q_i)$ and $\phi_Q(k_j)$ available in hand, the entry of $\hat{S}$ for standard Nyström approximation is calculated as,

$$\hat{S}_{ij} = \phi_K(q_i)^T \phi_Q(k_j), \forall i = 1,\dots,n, j = 1,\dots,n \tag{2}$$

To derive the explicit Nyström form, $\hat{S}$, of the softmax matrix, we assume that A_S is non-singular first to guarantee that the above expression to define ϕ_K and ϕ_Q is meaningful. We will shortly relax this assumption to achieve a general form.

When A_S is non-singular,

$$\begin{aligned}\hat{S}_{ij} &= \phi_K(q_i)^T \phi_Q(k_j) \\ &= [\mathcal{K}_K(q_i)]_{1\times m} V_{m\times m} \Lambda^{-1}_{m\times m} U^T_{m\times m} [\mathcal{K}_Q(k_j)]_{m\times 1}.\end{aligned}$$

Let $W_m = V_{m\times m}\Lambda^{-1}_{m\times m}U^T_{m\times m}$. Recall that a SVD of A_S is $U_{m\times m}\Lambda_{m\times m}V^T_{m\times m}$, and so, $W_m A_S = I_{m\times m}$. Therefore,

$$\hat{S}_{ij} = [\mathcal{K}_K(q_i)]_{1\times m} A_S^{-1} [\mathcal{K}_Q(k_j)]_{m\times 1} \tag{3}$$

Without requiring that A_S is non-singular, we can rewrite (3) as

$$\hat{S}_{ij} = [\mathcal{K}_K(q_i)^T]_{1\times m} A_S^{+} [\mathcal{K}_Q(k_j)]_{m\times 1}, \tag{4}$$

where A_S^+ is a Moore-Penrose pseudoinverse of A_S. So,

$$\hat{S}_{ij} = \left[\text{softmax}\left(\frac{q_i K^T}{\sqrt{d_q}}\right)\right]_{1\times m} A_S^{+} \left[\text{softmax}\left(\frac{Qk_j^T}{\sqrt{d_q}}\right)\right]_{m\times 1},$$

for $i, j = \{1,\dots,n\}$. The Nyström form of the softmax matrix, $S = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)$ is thus approximated as,

$$\hat{S} = \left[\text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)\right]_{n\times m} A_S^{+} \left[\text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)\right]_{m\times n}$$

where $[\cdot]_{n\times m}$ refers to taking m columns from $n\times n$ matrix and $[\cdot]_{m\times n}$ refers to taking m rows from $n\times n$ matrix.

Proofs of Theorems

This section details all our proofs of Lemma 1–2 and add Proposition 1. To keep the derivations succinct, we use $\mathscr{S}(\cdot)$ to denote softmax$(\cdot)$.

Lemma 1. *For $A_S \in \boldsymbol{R}^{m\times m}$, the sequence $\{Z_j\}_{j=0}^{j=\infty}$ generated by (Razavi et al. 2014),*

$$Z_{j+1} = \frac{1}{4} Z_j \left(13I - A_S Z_j (15I - A_S Z_j)(7I - A_S Z_j)\right) \tag{5}$$

converges to a Moore-Penrose inverse A_S^+ in the third-order with the initial approximation Z_0 satisfying $||A_S A_S^+ - A_S Z_0|| < 1$.

Proof. Let $\bar{E}_j = I - A_S Z_j$, then

$$\begin{aligned}
\bar{E}_{j+1} =& I - A_S Z_{j+1} \\
=& I - A_S \left(\frac{1}{4} Z_j (13I - A_S Z_j (15I - A_S Z_j)(7I - A_S Z_j) \right) \\
=& I - \frac{1}{4} A_S Z_j \left(13I - A_S Z_j (15I - A_S Z_j)(7I - A_S Z_j)\right) \\
=& \frac{1}{4} \left(4I - A_S Z_j (13I - A_S Z_j (15I - A_S Z_j)(7I - A_S Z_j))\right) \\
=& \frac{1}{4} \left(4I - A_S Z_j (13I - 15 A_S Z_j + 7 (A_S Z_j)^2 - (A_S Z_j)^3)\right) \\
=& \frac{1}{4} (4I - A_S Z_j)(I - A_S Z_j)^3 \\
=& \frac{1}{4} (3\bar{E}_j^3 + \bar{E}_j^4)
\end{aligned}$$

Therefore we get,

$$\bar{E}_{j+1} = \frac{1}{4} (3\bar{E}_j^3 + \bar{E}_j^4) \tag{6}$$

Let $E_j = Z_j - A_S^+$, we have,

$$\begin{aligned}
A_S E_{j+1} =& A_S Z_{j+1} - A_S A_S^+ \\
=& A_S Z_{j+1} - I + I - A_S A_S^+ \\
=& - \bar{E}_{j+1} + I - A_S A_S^+ \\
=& - \frac{1}{4} (3\bar{E}_j^3 + \bar{E}_j^4) + I - A_S A_S^+ \\
=& \frac{1}{4} (-3\bar{E}_j^3 + 3(I - A_S A_S^+)) + \frac{1}{4} (-\bar{E}_j^4 + (I - A_S A_S^+))
\end{aligned}$$

From the definition of Moore-Penrose pseudoinverse A_S^+, we have

$$(I - A_S A_S^+)^j = I - A_S A_S^+$$

and

$$(I - A_S A_S^+) A_S E_j = 0, \; j = 1, 2, \ldots n.$$

Therefore, $-\bar{E}_j^3 + (I - A_S A_S^+)$ is

$$\begin{aligned}
& = -(I - A_S Z_j)^3 + (I - A_S A_S^+) \\
& = -(I - A_S A_S^+ + A_S A_S^+ - A_S Z_j)^3 + (I - A_S A_S^+) \\
& = -((I - A_S A_S^+)^3 + 3(I - A_S A_S^+)^2 (A_S A_S^+ - A_S Z_j) \\
& + 3(I - A_S A_S^+)(A_S A_S^+ - A_S Z_j)^2 + (A_S A_S^+ - A_S Z_j)^3) \\
& + (I - A_S A_S^+) \\
& = -((I - A_S A_S^+) - 3(I - A_S A_S^+) A_S E_j \\
& + 3(I - A_S A_S^+)(A_S E_j)^2 - (A_S E_j)^3 + (I - A_S A_S^+) \\
& = -(A_S E_j)^3.
\end{aligned}$$

We can similarly show that

$$-\bar{E}_j^4 + (I - A_S A_S^+) = (A_S E_j)^4. \tag{7}$$

Then we have,

$$A_S E_{j+1} = \frac{1}{4} (-3(A_S E_j)^3 + (A_S E_j)^4) \tag{8}$$

Thus we obtain,

$$||A_S E_{j+1}|| \leq \frac{1}{4} (3||A_S E_j||^3 + ||A_S E_j||^4)$$

With the assumption on the initial approximation, $||A_S A_S^+ - A_S Z_0|| < 1$, we have $||A_S E_0|| < 1$. Therefore, based on (8), we obtain $||A_S E_j|| < 1$. Thus we have,

$$||A_S E_{j+1}|| \leq \frac{1}{4} (3||A_S E_j||^3 + ||A_S E_j||^3) \leq ||A_S E_j||^3$$

For $E_{j+1} = Z_{j+1} - A_S^+$, we have

$$\begin{aligned}
||E_{j+1}|| &= ||A_S^+ A_S Z_{j+1} - A_S^+ A_S A_S^+|| \\
&\leq ||A_S^+||\, ||A_S Z_{j+1} - A_S A_S^+|| \\
&= ||A_S^+||\, ||A_S E_{j+1}|| \\
&\leq ||A_S^+||\, ||A_S E_j||^3
\end{aligned}$$

As we have $||A_S E_j|| < 1$, we get $||Z_j - A_S^+|| \to 0$ in third order as $j \to +\infty$. It thus concludes the proof. □

Lemma 2. *Given the input data set $Q = \{q_i\}_{i=1}^n$ and $K = \{k_i\}_{i=1}^n$, and the corresponding landmark point set $\tilde{Q} = \{\tilde{q}_j\}_{j=1}^m$ and $\tilde{K}_j = \{\tilde{k}\}_{j=1}^m$. Using* (10)*, the Nyström approximate self-attention converges to true self-attention if there exist landmarks points $\tilde{q}_p$ and $\tilde{k}_t$ such that $\tilde{q}_p = q_i$ and $\tilde{k}_t = k_j$, $\forall i = 1, \ldots, n, j = 1, \ldots, n$.*

Proof. For all $i = 1, \ldots, n, j = 1, \ldots, n$, there exist landmarks points $\tilde{q}_p$ and $\tilde{k}_t$ such that $\tilde{q}_p = q_i$ and $\tilde{k}_t = k_j$. It means that we can obtain a landmark matrix $\tilde{Q} = Q$ and $\tilde{K} = K$. Using the Nyström approximation,

$$\hat{S} = \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) Z_l \, \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) \tag{9}$$

where Z_l is generated by (5). Then, the error between approximate self-attention and the true self-attention with the ℓ_∞ norm is,

$$\begin{aligned}
& ||SV - \hat{S}V||_\infty \\
& = \|\mathscr{S}(\frac{QK^T}{\sqrt{d_q}})V - \mathscr{S}(\frac{Q\tilde{K}^T}{\sqrt{d_q}}) Z_l \mathscr{S}(\frac{\tilde{Q}K^T}{\sqrt{d_q}})V\|_\infty \\
& = \|\mathscr{S}(\frac{QK^T}{\sqrt{d_q}}) \mathscr{S}(\frac{QK^T}{\sqrt{d_q}})^+ \mathscr{S}(\frac{QK^T}{\sqrt{d_q}})V \\
& - \mathscr{S}(\frac{Q\tilde{K}^T}{\sqrt{d_q}}) Z_l \mathscr{S}(\frac{\tilde{Q}K^T}{\sqrt{d_q}} V\|_\infty \\
& = \|\mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right) (\mathscr{S}(\frac{QK^T}{\sqrt{d_q}})^+ - Z_l) \mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right) V\|_\infty \\
& \leq \|\mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right)^+ - Z_l\|_\infty \|V\|_\infty
\end{aligned}$$

The last inequality holds since $\|\mathscr{S}(\frac{QK^T}{\sqrt{d_q}})\|_\infty = 1$. Note that Lemma 2 shows that the sequence Z_l generated by Eq. 5 converges to the Moore-Penrose inverse ($Z_l \to \mathscr{S}(\frac{QK^T}{\sqrt{d_q}})^+$). Therefore, the approximate self-attention will converge to the true self-attention. This concludes the proof. □

Proposition 1. *The error of our Nyström approximation, i.e., the difference between the approximate self-attention,*

$$\hat{S}V = \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) Z^\star \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) V, \tag{10}$$

and the true self-attention,

$$SV = \mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right) V, \tag{11}$$

is bounded by

$$E \leq (1 + ||A_S^+||_\infty + ||A_S^+ - Z^\star||_\infty)||V||_\infty \tag{12}$$

where E denotes the error of the approximation in ℓ_∞ norm,

$$\|\mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right) V - \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) Z^\star \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) V\|_\infty.$$

Q denotes the input query, K is the input key, and V is the input value. $\tilde{Q}$ and $\tilde{K}$ are corresponding landmark point matrices of Q and K, respectively. And $Z^\star$ is the approximate of the Moore-Penrose inverse of A_S.

Proof. With the definition of E, we have

$$\begin{aligned}
E &\leq \|\mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right) - \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) Z^\star \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right)\|_\infty \|V\|_\infty \\
&= ||\mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right) - \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) \mathscr{S}\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_q}}\right)^+ \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) \\
&+ \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) \mathscr{S}\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_q}}\right)^+ \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) \\
&- \mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) Z^\star \mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) ||_\infty ||V||_\infty \\
&\leq (\|\mathscr{S}\left(\frac{QK^T}{\sqrt{d_q}}\right)\|_\infty \\
&+ \|\mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right)\|_\infty \|A_S^+\|_\infty \|\mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right)\|_\infty \\
&+ \|\mathscr{S}\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right)\|_\infty \|A_S^+ - Z^\star\|_\infty \|\mathscr{S}\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right)\|_\infty) \|V\|_\infty \\
&= (1 + \|A_S^+\|_\infty + \|A_S^+ - Z^\star\|_\infty)\|V\|_\infty
\end{aligned}$$

which concludes the proof. □

Implementation Details

Implementation details. We describe pre-training details for our experiments on BookCorpus plus English Wikipedia. Our model is pretrained with the masked-language-modeling (MLM) and sentence-order-prediction (SOP) objectives (Lan et al. 2020) on BookCorpus plus English Wikipedia. We use a batch size of 256, optimizer Adam with learning rate 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, learning rate warmup over the first 10,000 steps, and linear learning rate decay to update our model. For BERT-small comparison, We train all models in 0.1M steps. For BERT-base comparison, our baseline was trained with 0.5M steps and our model was trained with $\sim$ 0.25M steps, initialized from pretrained BERT-base.

Model	SST-2	MRPC	QNLI	QQP	MNLI m/mm	IMDB
BERT-base	90.0	88.4	90.3	87.3	82.4/82.4	93.3
Nyströmformer (512)	91.4	88.1	88.7	86.3	80.9/82.2	93.2
Nyströmformer (1024)	91.4	87.5	88.7	86.3	80.9/81.4	93.0

Table 1: Dev set results on benchmark natural language understanding tasks. We report F1 score for QQP and accuracy for others. Our Nyströmformer performs competitively with BERT-base and the Nyströmformer pretrained with longer sequence length 1024 has fairly similar results to the one pretrained with shorter length 512 on benchmark natural language understanding task.

Further Experimental Results

To complement our main results, we present additional experiments on landmark selection, iterative approximation of the pseudoinverse, inference using longer sequences.

(A) Landmark selection. We now provide an ablation study of the landmark selection step in our model. The goal is to evaluate if the selected landmarks are sufficient to reconstruct the self-attention. **(a) Setup.** We experiment with various methods for computing landmark points. We compare our segmented means with uniform random sampling and K-means. To achieve comparisons on real data, we extract query Q, key K, value V of 6 different heads from a trained BERT-base model. **(b) Findings.** Fig 1 visualizes the approximate self-attention with different landmark selection schemes. The results show that our Segment-means with 64 landmark points outperforms uniform random sampling and compares favorably to K-means. Importantly, our method only has a linear runtime footprint, and thus is more efficient than the iterative procedure used by K-means. Further, we study the choice of numbers of landmark points. Similarly, we extract query, key, value of 6 different heads from trained BERT-base model. Fig. 2 shows the experimental results when using $16, 64, 256$ landmark points and compares them to ground-truth self-attention. Not surprisingly, the more landmarks the better the approximation. With 256 landmark points, the approximation to softmax matrix is more accurate, yet requires more time to compute the approximate Moore-Penrose inverse. Using 16 landmark points is more efficient yet with increased approximation error. To balance the efficiency and approximation accuracy, we use 64 landmarks for our model.

(B) Iterative approximation of pseudoinverse. We conduction further experiments to verify the quality of our approximate pseudoinverse. **(a) Setup.** We compare our approximation to pseudoinverse computed using `numpy.linalg.pinv`. Similarly, we extract query, key, and value from trained BERT-base for comparison. **(b) Findings.** We find that the iterative method achieves a good approximation of the ground truth only with 6 iterations in Fig. 3. Fig. 4 further visualizes the ground truth and our approximate pseudoinverse from 6 different heads.

(C) Longer sequences. We also experiment with inference using longer sequences. **(a) Setup.** We test our model with longer input length ($n = 1024$) after pretraining. To fit the longer sequence input, we increase the position embedding dimension from 512 to 1024. Following (Wang et al. 2020), we train our model from trained model with input sequence length 512. We use a batch size of 128, optimizer Adam with learning rate 1e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, and linear learning rate decay to update our longer model. The longer model is trained with 40K steps. **(b) Findings.** Fig. 5 plots the validation accuracy of input sequence length 1024 for our model. We observe that when we increase the input sequence length, the final validation MLM accuracy will remain the same yet SOP accuracy will improve. For MLM accuracy, longer model is performing similarly as standard model. This is because MLM task is a local prediction task and thus is unlikely to benefit from a longer sequence. For SOP, longer input sequence contains more information for the task, and thus can improve the performance by 2%. These results further justify the necessity of our model, which can enable the training and inference on longer sequences. Furthermore, We finetune our pretrained model with longer sequence 1024 on GLUE benchmark datasets and IMDB reviews respectively and report its final performance. While downstreaming tasks do not exceed the maximum input sequence length 512, the results remain almost identical as n = 512 in Table 1. These results further indicates that our model is able to scale linearly with input length.

References

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Razavi, M. K.; Kerayechian, A.; Gachpazan, M.; and Shateyi, S. 2014. A new iterative method for finding approximate inverses of complex matrices. In *Abstract and Applied Analysis*.

Wang, S.; Li, B.; Khabsa, M.; Fang, H.; and Ma, H. 2020. Linformer: Self-Attention with Linear Complexity. *arXiv preprint arXiv:2006.04768* .

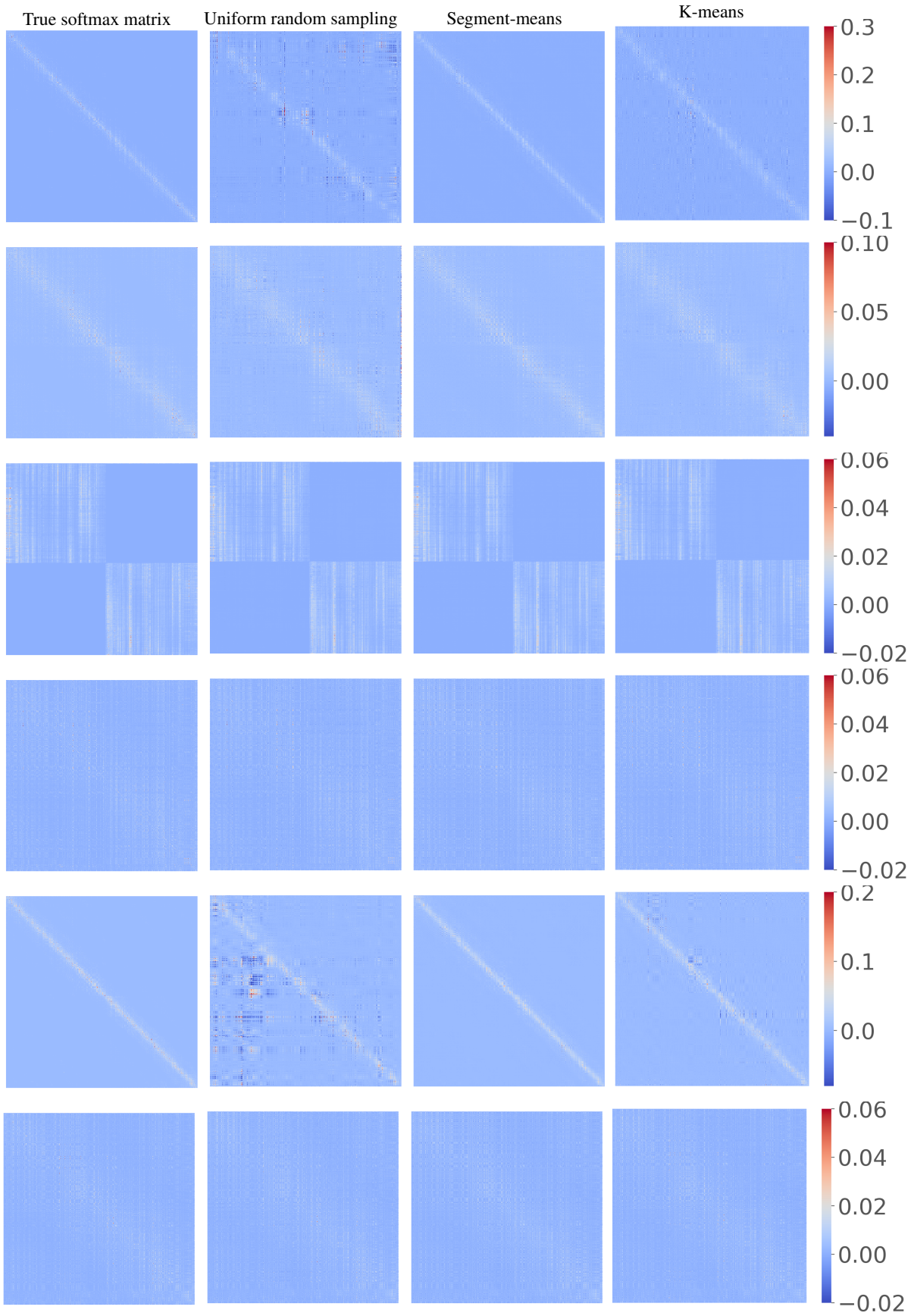

Figure 1: Landmarks selection comparison results. The most left column is ground truth softmax matrix of 6 different self-attention heads. Right columns are approximate softmax by using uniform random sampling, Segment-means and K-means to select landmark points. Segment-means performs favorably with k-means.

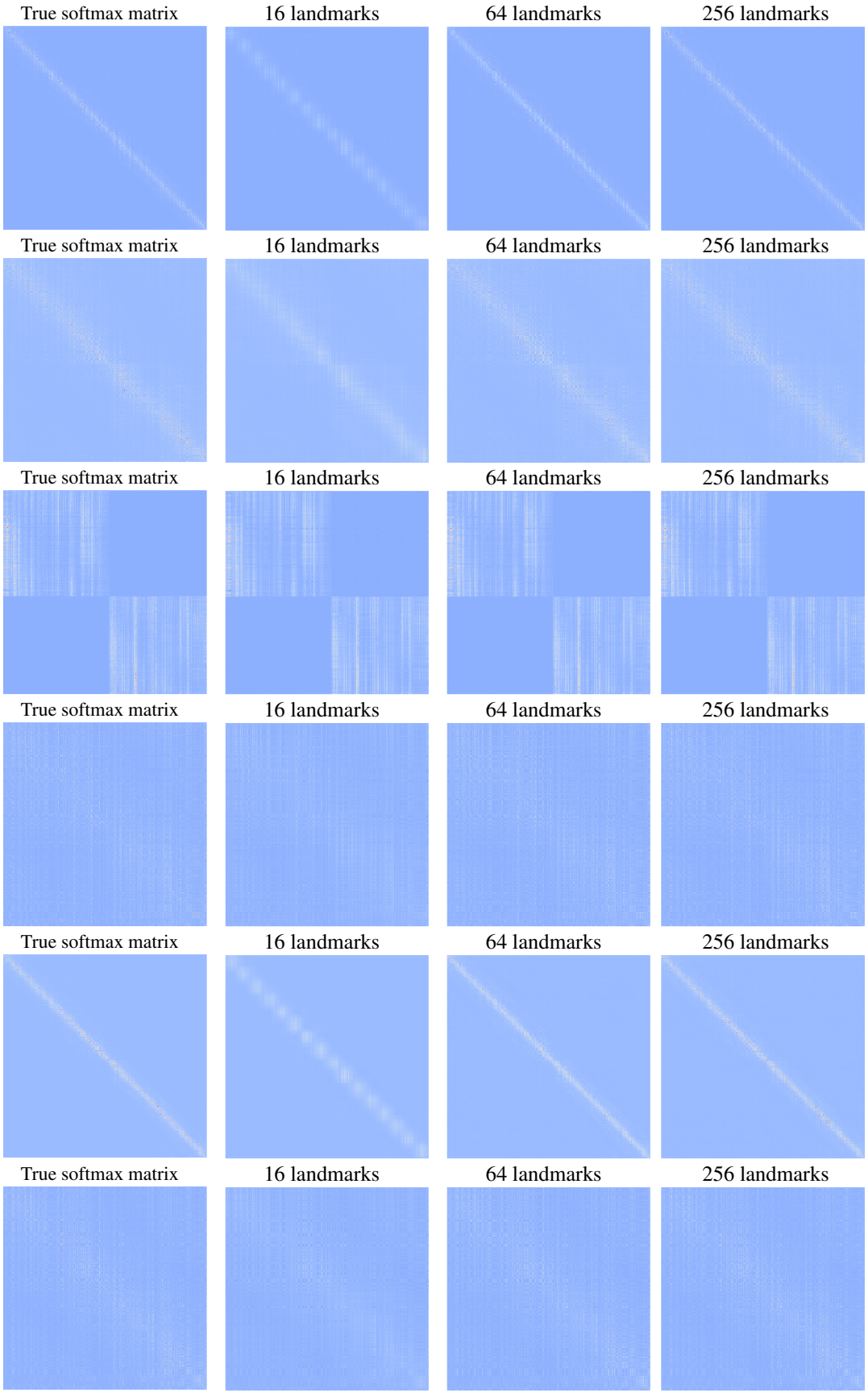

Figure 2: Different number of landmarks selected to approximate softmax matrix. The most left column is ground truth softmax matrix of 6 different self-attention heads. Right columns are approximate softmax with 16, 64, 256 landmark points, selected by Segment-means. Using more landmark points leads to more accurate approximation.

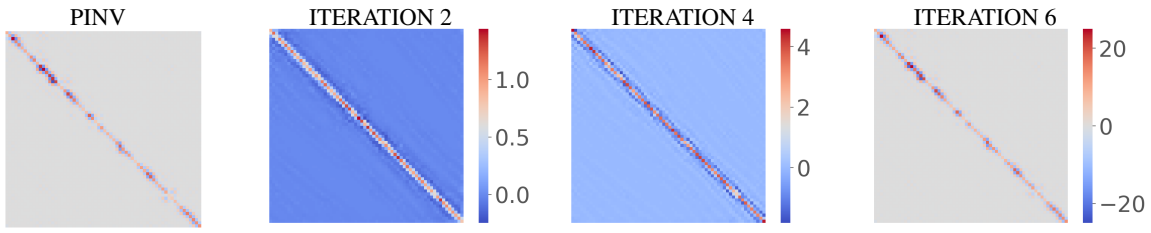

Figure 3: Iterative method for approximating true Moore-Penrose pseuduinverse. Left is the true pseudoinverse by *numpy.linalg.pinv*. Right is its approximation by Eq. 5 with 2, 4, 6 iterations. Running 6 iterations achieves a good approximation of the pseudoinverse.

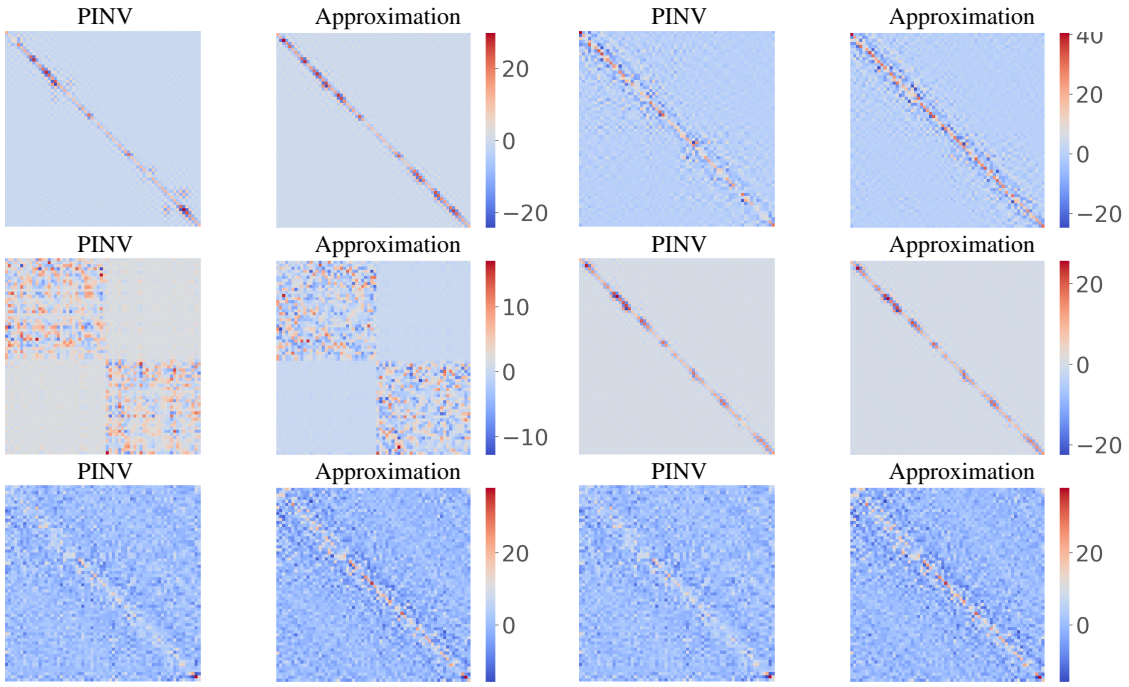

Figure 4: Approximating true Moore-Penrose pseudoinverse results. Odd columns are the true pseudoinverse from 6 different self-attention heads, computed by *numpy.linalg.pinv* . Even columns are their corresponding approximation by Eq. 5 with 6 iterations.

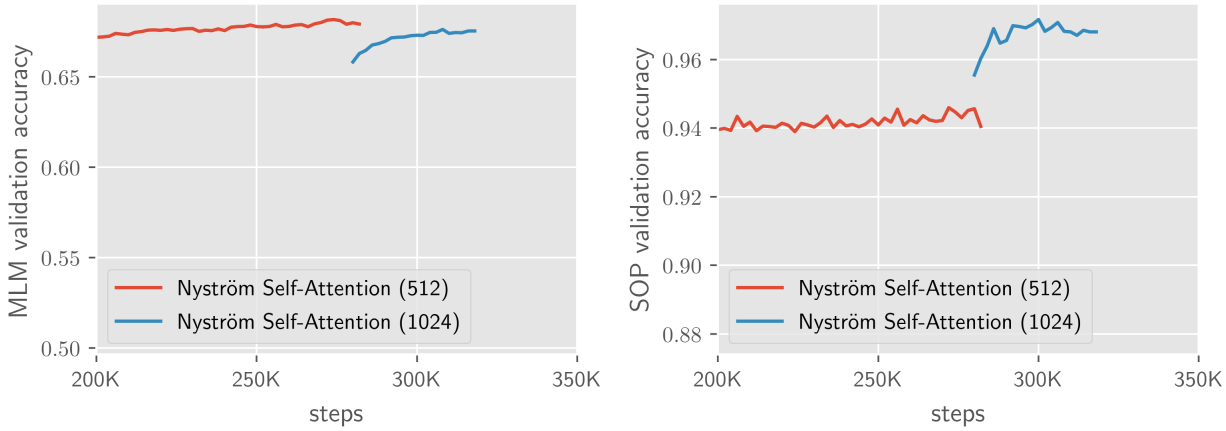

Figure 5: Results on MLM and SOP with longer sequence input length 1024. We report MLM and SOP validation accuracy. Our Nyströmformer (1024) is trained with additional ∼40K steps on input sequence with length 1024 after training ∼ 0.25 M steps on input sequence with length 512. Training on longer input sequence remains the same MLM performance as MLM is a local prediction task, while our model increases the SOP performance by 2% accuracy as SOP task relies on longer context.