IEOR E6617 Research Project Report: Adapting Linear Attention to Query/Key Distributions

Xingchen (Estella) Ye
xy2527@columbia.edu
Dec 11, 2023
Project GitHub Link: https://github.com/xcye515/high-dim-ml-f23

This research focuses on adapting linear attention mechanisms to approximated query and key vectors in high-dimensional spaces. Methods are proposed to approximate the means (q_0, k_0) and covariance matrices (Σ_q, Σ_k) of queries and keys, under the assumption that sufficient samples can be drawn from the true distributions.

Then, we explore methods to linearize the softmax attention kernel, introducing two random feature mappings $\phi^{sin/cos}(x, w)$ and $\phi(x, w)$, proposed in the works of Rahimi & Recht (2007) and Choromanski et al. (2021). Our research includes an analysis of the linearity of these methods and their unbiased nature. Additionally, we incorporate the squared norm approximation technique to efficiently compute the norms of queries and keys.

Through series of experiments, we examine the impact of varying sample sizes on the attention mechanism. Then, we demonstrate how the specific underlying distributions, $\mathcal{N}(q_0, \Sigma_1), \mathcal{N}(k_0, \Sigma_k)$, influence the behavior of the attention approximation. The results are visualized through a series of figures, providing valuable insights into the behavior of attention approximations under different distribution constructions.

I. Approximation of $q_0, k_0, \Sigma_0, \Sigma_k$

We assume that the queries and keys $q \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$ can be efficiently sampled from the distributions. It is advisable to draw at least $N > d^2$ samples to reduce the curse of dimensionality. Given the sets of samples $S_q = \{q_1, q_2, \ldots, q_N\}$ and $S_k = \{k_1, k_2, \ldots, k_N\}$, which are drawn from the respective distributions, the sample means can be computed using the following formulas:

$$\hat{q}_0 = \frac{1}{N}\sum_{i=1}^{N} q_i$$
$$\hat{k}_0 = \frac{1}{N}\sum_{i=1}^{N} k_i \tag{1}$$

The estimations for the covariance matrices are calculated as:

$$\hat{\Sigma}_q = \frac{1}{N-1}\sum_{i=1}^{N}(q_i - \hat{q}_0)(q_i - \hat{q}_0)^\mathsf{T}$$
$$\hat{\Sigma}_k = \frac{1}{N-1}\sum_{i=1}^{N}(k_i - \hat{k}_0)(k_i - \hat{k}_0)^\mathsf{T} \tag{2}$$

II. Approximation of the attention matrix $\mathbf{A}$

Denote the matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{\mathbf{N}\times\mathbf{D}}$ for queries, keys and values. The softmax attention can be computed using the formula:

$$A_{\text{softmax}}(q_n, \mathbf{K}, \mathbf{V}) = \sum_{n=1}^{N} \frac{\exp(q_n^\mathsf{T} k_n)}{\sum_{n'=1}^{N}\exp(q_n^\mathsf{T} k_{n'})} v_n^\mathsf{T} \tag{3}$$

where the kernel function $\mathbf{K}$ is defined as:

$$K(x, y) = \exp(x^\mathsf{T} y) \tag{4}$$

Rahimi & Recht (2007) propose that if $\mathbf{K}$ is a positive semidefinite kernel, then by Mercer's theorem, there exists a feature map $\phi : \mathbb{R}^D \to \mathbb{R}^V$ such that:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_v \tag{5}$$

They further discuss in the paper that this inner product in $\mathbb{R}^V$ can be approximated by a randomized feature map $z : \mathbb{R}^D \to \mathbb{R}^R$ by Bochner's theorem. The formal derivation can be found in both [1] and [2].

$$K(x, y) \approx z(x)^\mathsf{T} z(y) \tag{6}$$

Choromanski et al. (2021) and Peng et al. (2021) propose methods to linearize the exponential kernels by applying a random feature transformation to the kernel function. According to Bochner's theorem, the kernel can be reformulated as follows:

$$K(x, y) = \mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}[\phi(x, w)^\mathsf{T}\phi(y, w)] \tag{7}$$

where the feature mapping function $\phi : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^l$ transforms the input vector into a lower-dimensional vector, and the weight w follows a spherical Gaussian distribution, $w \sim \mathcal{N}(0, \mathbf{I})$. Here we give two feature mappings.

1) $\phi^{sin/cos(x,w)}$ Feature Mapping

Rahimi & Recht (2008) and Peng et al. (2021) introduce the feature mapping $\phi^{\text{sin/cos}}$, which approximates the Gaussian kernel K_{softmax}. The mapping is defined as:

$$\phi^{\text{sin/cos}}(x, w) = \exp\left(\frac{\|x\|^2}{2}\right)\begin{bmatrix}\cos(w^\mathsf{T} x)\\ \sin(w^\mathsf{T} x)\end{bmatrix} \tag{8}$$

2) $\phi(x, w)$ Randomized Mapping

In a different approach, Choromanski et al. (2021) propose a scalar-valued, positive, randomized mapping $\phi(x, w)$, defined by:

$$\phi(x, w) = \exp\left(w^\mathsf{T} x - \frac{1}{2}\|x\|^2\right) \tag{9}$$

This mapping emphasizes the importance of non-linear transformations in kernel approximation.

In this report, we focus on the $\phi(x, w)$ randomized mapping, and we aim to show the linearity of this approach in the following section.

III. Linearity of Kernel Approximation and Analysis on Biases

In the work of Choromanski et al. (2021), the expectation of $K(x, y)$ is approximated by the following equation:

$$A_{\text{softmax}}(q_n, K, V) \approx \frac{\sum_{s=1}^{S} \phi(q_n, w_s)^\mathsf{T} \sum_{n=1}^{N} \phi(k_n, w_s) v_n^\mathsf{T}}{\sum_{s=1}^{S} \phi(q_n, w_s)^\mathsf{T} \sum_{n'=1}^{N} \phi(k_{n'}, w_s)} = RFA(q_n, K, V) \tag{10}$$

Zheng et al. (2022) states that this formulation commonly employs Importance Sampling (IS) method for computation: In scenarios where sampling directly from $p(w)$ and computing $E_{p(w)}[f(w)]$ is challenging, samples are drawn from a proposal distribution $q(w)$, using the following estimation for the original expectation:

$$E_{p(w)}[f(w)] \approx \frac{1}{S} \sum_{s=1}^{S} \frac{p(w_s)}{q(w_s)} f(w_s) \tag{11}$$

Zheng et al. (2022) further argues that this estimation approach does not satisfy the conditions for using IS, because the distribution $p(w)$ cannot be evaluated explicitly. They suggest that Self-Normalized Importance Sampling (SNIS) is a more appropriate method. However, SNIS introduces an inherent bias by normalizing the ratio $p(w)/q(w)$, indicating that this method is intrinsically biased.

To reduce the bias, Zheng et al. (2022) propose an alternative method for estimating the expectation of kernel. In this approach, S samples of w are drawn from the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. The estimation is given by:

$$A_{\text{softmax}}(q_n, K, V) \approx \frac{1}{S} \sum_{s=1}^{S} \frac{\phi(q_n, w_s)^\mathsf{T} \sum_{n=1}^{N} \phi(k_n, w_s) v_n^\mathsf{T}}{\phi(q_n, w_s)^\mathsf{T} \sum_{n'=1}^{N} \phi(k_{n'}, w_s)} = RA(q_n, K, V) \tag{12}$$

This method simplifies the sampling process by drawing the weights directly from $p(w) = \mathcal{N}(0, \mathbf{I})$. This approach differs from the RFA method by computing attention-like aggregations.

IV. Approximation of the Norm

The computation of norm values for a set of queries/keys is essential in both feature mappings presented in Equations 8 and 9. The estimated means $\hat{q_0}, \hat{k_0}$ and the covariance matrices $\hat{\Sigma_q}, \hat{\Sigma_k}$ can be used to approximate these norms. The Squared Norm Approximation is defined as follows:

$$||x||^2 \approx ||\mu_x||^2 + \text{Tr}(\Sigma_x) \tag{13}$$

Here, μ_x is the mean of x, and $\text{Tr}(\Sigma_x)$, the trace of Σ_x, is the sum of the diagonal elements (the variances) in the covariance matrix. The term $||\mu_x||^2$ calculates the squared norm of the mean vector, representing the central point of the distribution. $\text{Tr}(\Sigma_x)$ adds the total variance of the vector elements, which is the spread of x around its mean.

By utilizing the squared norm approximation, we can simplify the calculations for each feature mapping. Instead of computing the norm for each query/key at every position, we calculate $q' = ||\hat{q_0}||^2$ and $k' = ||\hat{k_0}||^2$ just once.

Applying the approximation method to Equation 8:

$$z^{\text{sin/cos}}(q_n, w) = \exp\left(\frac{q'}{2}\right) \begin{bmatrix} \cos(w^\mathsf{T} q_n) \\ \sin(w^\mathsf{T} q_n) \end{bmatrix}$$
$$z^{\text{sin/cos}}(k_n, w) = \exp\left(\frac{k'}{2}\right) \begin{bmatrix} \cos(w^\mathsf{T} k_n) \\ \sin(w^\mathsf{T} k_n) \end{bmatrix} \tag{14}$$

Applying the approximation method to Equation 9:

$$z(q_n, w) = \exp\left(w^\mathsf{T} q_n - \frac{1}{2} q'\right)$$
$$z(k_n, w) = \exp\left(w^\mathsf{T} k_n - \frac{1}{2} k'\right) \tag{15}$$

V. Experiment

The experiment involves several steps:

1) generate query-key pairs and approximate the sample means and sample covariances.
2) implement the random feature map in Equation 9 to approximate the Guassian attention. Experiments are conducted to compare the attention generated by varying sample sizes of weights R under different distribution configurations.

A. Approximation of $q_0, k_0, \Sigma_q, \Sigma_k$

In our experiment, we begin with the assumption that we know the distributions for queries and keys. To initiate this process, we first draw the true means q_0 and k_0 from a uniform distribution $\mathcal{U}(\mathbf{0}, \mathbf{1})$ of D-dimensional vectors. Subsequently, we construct the true covariance matrices, Σ_q and Σ_k. These matrices, being of dimensions $D \times D$, are required to be positive semi-definite. We could first generate two matrices $A, B \sim \mathcal{U}(0, 1)^{D \times D}$, and then let $\Sigma_q = AA^T$ and $\Sigma_k = BB^T$.

With the true means and covariance matrices for queries and keys, we proceed to sample from these distributions with a sample size of N. Once the samples are obtained, we use Equations 1 and 2 to calculate the sample means, $\hat{q_0}$ and $\hat{k_0}$, as well as the sample covariance, $\hat{\Sigma_q}$ and $\hat{\Sigma_k}$. These computed values are employed to approximate the norms and $\phi(q, w)$ and $\phi(k, w)$.

Across the experiments, we standardize the query/key sample size at $N = 1000$ and feature dimension $D = 10$ for consistent comparison.

B. Approximation of attention matrix A

To compare the effects of the sample size R on the attention, we would need obtain $z(q, w), z(k, w)$ for different sets of w. We conduct the experiments on $R = 1, 10, 100, 1000, 10000$.

First, we apply the randomized mapping on $\hat{q_0}, \hat{k_0}$ (Equation 15). Note that the value of R would affect the linear attention approximation significantly. Then, in the original

randomized feature map, the calculation of Euclidean norm for each query and key is required. With the approximated values $\hat{q_0}, \hat{k_0}, \hat{\Sigma_q}, \hat{\Sigma_k}$, squared norm approximation allows us to compute the norm only for once (q', k' in Equation 15), effectively reducing the computational costs.

After $\phi(q, w)$ and $\phi(k, w)$ are obtained, we could approximate the brute-force attention matrix A by Equation 6.

C. Results

The visualization of the approximated linear attention is presented in Figure 1, which was constructed based on the methods described in Subsection A. Note that, the true means and covariance matrices were randomly generated from uniform distributions.

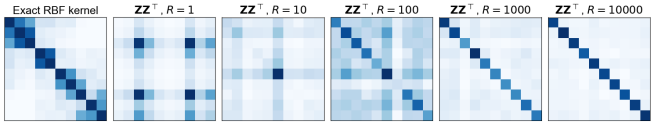

Fig. 1. Attention using samples from key/query distribution: $q_0, k_0 \sim \mathbb{U}(\mathbf{0}, \mathbf{1}), \Sigma_q, \Sigma_k \sim \mathbb{U}(0, 1)^{D \times D}$

Our experiments further explore the impact of various distributions $\mathcal{N}(q_0, \Sigma_q)$ and $\mathcal{N}(k_0, \Sigma_k)$ on the attention approximation. Specifically, by setting q_0 and k_0 as constant vectors $\mathbb{1}$ and $\mathbb{3}$, distinct outcomes are observed and illustrated in the following figures.

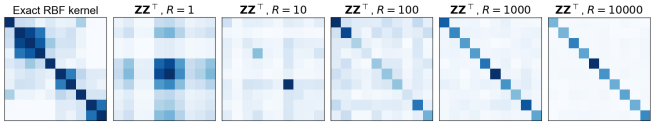

Fig. 2. Attention using samples from key/query distribution: $q_0 = \mathbf{1}, k_0 = \mathbf{1}, \Sigma_q, \Sigma_k \sim \mathbb{U}(0, 1)^{D \times D}$

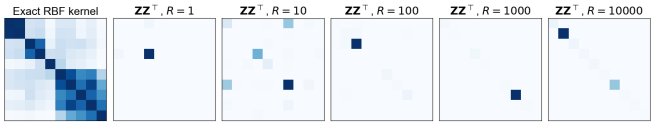

Fig. 3. Attention using samples from key/query distribution: $q_0 = \mathbf{3}, k_0 = \mathbf{3}, \Sigma_q, \Sigma_k \sim \mathbb{U}(0, 1)^{D \times D}$

Additionally, we experiment with setting Σ_q and Σ_k to the product of AA^T, where A is a matrix with all elements 1 and 3. The resulting visualizations are given as follows:

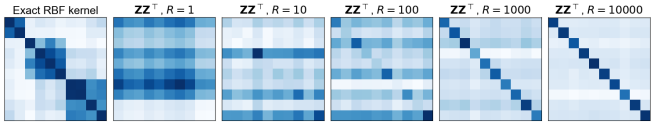

Fig. 4. Attention using samples from key/query distribution: $q_0, k_0 \sim \mathbb{U}(\mathbf{0}, \mathbf{1}), \Sigma_q, \Sigma_k = AA^\mathsf{T}$ where $A = 1^{D \times D}$

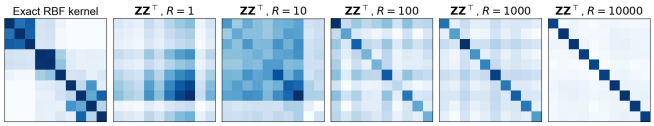

Fig. 5. Attention using samples from key/query distribution: $q_0, k_0 \sim \mathbb{U}(\mathbf{0}, \mathbf{1}), \Sigma_q, \Sigma_k = AA^\mathsf{T}$ where $A = 3^{D \times D}$

VI. Conclusion

This study demonstrates that the sample size of weights, R, along with the chosen configurations of the true distributions $\mathcal{N}(q_0, \Sigma_q), \mathcal{N}(k_0, \Sigma_k)$, significantly influence the precision of attention approximation. Notably, it is observed that an increase in R generally leads to better approximation. However, an exception is noted in the case of Figure 3, where $q_0, k_0 = \mathbf{3}$ and $R = 10000$ fails to approximate the attention. These findings demonstrate the importance of selecting the appropriate approximation method and the sample size R, while maintaining efficient computational costs. Future study could experiment with higher feature-space dimension D, to further understand these linear approximation techniques.

References

[1] Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc., 2008.

[2] Gundersen, G.. Random Fourier Features. Random fourier features. https://gregorygundersen.com/blog/2019/12/23/random-fourier-features/

[3] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L. and Belanger, D., Rethinking attention with performers, 2021.

[4] Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N., and Kong, L. Random feature attention. In International Conference on Learning Representations, 2021.

[5] Lin, Z., Chong, W., Lingpeng, K., Linear Complexity Randomized Self-attention Mechanism, 2021.