Supplemental Material: KVT: *k*-NN Attention for Boosting Vision Transformers

Pichao Wang*, Xue Wang*, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin

Alibaba Group
{pichao.wang, xue.w, fan.w, ming.l, shuning.csn, lihao.lh, jinrong.jr}@alibaba-inc.com

1 Differences with the arXiv paper: Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection (EST)

Similarities: Our method part is similar to EST in terms of the calculation of top-k.
Differences:

- Our paper is focused not only on the methodology part, but also the deep understanding. There are many variants of Transformers in the NLP and vision community now, but few of them provide a deep and thorough analysis of their proposed methods. The proposed k-NN attention indeed happens to be similar to EST, which was arxived 2 years ago and we were not aware of it when conducting our research. In addition to applying the idea to transformers and conducting extensive experiments as EST did, we provide theoretical justifications about the idea, which we think is equally or more important than the method itself, and helps with a more fundamental understanding.
- The conclusion about how to select k is different. In EST, it is found that a small k is better (8 or 16), but we find a larger k, namely, $\geq \frac{1}{2}N$ is better (N is the sequence length).
- The motivations of these two papers are different: EST targets to get sparse attention maps while ours aims to distill noisy patches.
- Our paper focuses on vision transformers but EST focuses on NLP tasks, even though EST applied it to the image captioning task. Since late 2020, vision transformer backbones have become very popular, and k-NN attention deserves a deeper analysis. Therefore, we apply the k-NN attention on 11 different vision transformer backbones for empirical evaluations and find it simple and effective for vision transformer backbones.
- More analysis about the properties of k-NN attention in the context of vision transformer backbones are provided in our paper. Besides the k selection

* The first two authors contribute equally.

and convergence speed as EST presented, we also define several metrics to facilitate the analysis, e.g. layer-wise cosine similarity between tokens, layer-wise standard deviation of attention weights, ratio between the norms of residual activation and main branch, and nonlocality. We also compare it with temperature in $softmax$ and provide the visualizations.

In summary, our paper provides deeper understanding with comprehensive analysis of the k-NN attention for vision transformers, which provides well-grounded knowledge advancement.

2 Source codes of fast version *k*-NN attention in Pytorch

The source codes of fast version k-NN attention in Pytorch are shown in Algorithm 1, and we can see that the core codes of fast version k-NN attention is consisted of only four lines, and it can be easily imported to any architecture using fully-connected attention.

Algorithm 1 Codes of fast version k-NN attention in Pytorch.

```
class kNN-Attention(nn.Module):
    def __init__(self,dim,num_heads=8,qkv_bias=False,qk_scale=None,attn_drop
    =0.,proj_drop=0.,topk=100):
        super().__init__()
        self.num_heads=num_heads
        head_dim=dim//num_heads
        self.scale=qk_scale or head_dim**-0.5
        self.topk=topk

        self.qkv=nn.Linear(dim,dim*3,bias=qkv_bias)
        self.attn_drop=nn.Dropout(attn_drop)
        self.proj=nn.Linear(dim,dim)
        self.proj_drop=nn.Dropout(proj_drop)

    def forward(self,x):
        B,N,C=x.shape
        qkv=self.qkv(x).reshape(B,N,3,self.num_heads,C//self.num_heads).
    permute(2,0,3,1,4)
        q,k,v=qkv[0],qkv[1],qkv[2] #B,H,N,C
        attn=(q@k.transpose(-2,-1))*self.scale #B,H,N,N
        # the core code block
        mask=torch.zeros(B,self.num_heads,N,N,device=x.device,requires_grad=
    False)
        index=torch.topk(attn,k=self.topk,dim=-1,largest=True)[1]
        mask.scatter_(-1,index,1.)
        attn=torch.where(mask>0,attn,torch.full_like(attn,float(’-inf’)))
        # end of the core code block
        attn=torch.softmax(attn,dim=-1)
        attn=self.attn_drop(attn)
        x=(attn@v).transpose(1,2).reshape(B,N,C)
        x=self.proj(x)
        x=self.proj_drop(x)

        return x
```

3 Comparisons between slow version and fast version

We develop two versions of k-NN attention, one slow version and one fast version. The k-NN attention is exactly defined by slow version, but its speed is extremely slow, as for each query it needs to select different k keys and values, and this procedure is very slow. To speedup, we developed the CUDA version, but the speed is still slower than fast version. The fast version takes advantages of matrix multiplication and greatly speedup the computing. The speed comparisons on DeiT-Tiny using 8 V100 are illustrated in Table 1.

method	time per iteration (second)
slow version (pytorch)	8192
slow version (CUDA)	1.55
fast version (pytorch)	0.45

Table 1. The speed comparisons on DeiT-tiny for slow and fast version

4 Evaluations on CIFAR10 or CIFAR100.

As vision transformers are data-hungry, directly training vision transformer backbones from scratch on small-size datasets such as CIFAR10 or CIFAR100 would yield much worse performances compared with ConvNets. Following the paradigm and codes of the NIPS2021 paper "Efficient Training of Visual Transformers with Small Datasets", we briefly conducted experiments on CIFAR10 and CIFAR100 using Swin-T and T2T-ViT-14 with k-NN attention as shown in Table 2. Adding k-NN attention brings much larger performance gain in the scratch training (ST) due to its faster convergence speed, while the gain in the setting of ImageNet-1k pretraining and CIFAR finetuning (FT) is not as large.

Model	C10 (ST)	C100 (ST)	C10 (FT)	C100 (FT)
Swin-T	83.9	66.2	98.4	88.4
Swin-T→ k-NN Attn	84.5	67.1	98.6	88.7
T2T-VIT-14	87.6	68.0	98.5	87.7
T2T-VIT-14→ k-NN Attn	88.2	68.8	98.8	88.1

Table 2. Results on CIFAR10 and CIFAR100 (100 epochs).

5 Proof

Notations. Throughout this appendix, we denote x_i as i-th element of vector $\boldsymbol{x}$, $\boldsymbol{W}_{ij}$ as the element at i-th row and j-th column of matrix $\boldsymbol{W}$, and $\boldsymbol{W}_j$ as the

j-th row of matrix $\boldsymbol{W}$. Moreover, we denote $\boldsymbol{x}_i$ as the i-th patch (token) of the inputs with $\boldsymbol{x}_i = \boldsymbol{X}_i$.

Proof for Lemma 1 We first give the formal statement of Lemma 1.

Lemma 1 (Formal statement of Lemma 1). *Let $\hat{\boldsymbol{V}}_l^{knn}$ be the l-th row of the $\hat{\boldsymbol{V}}^{knn}$ and $Var_{\boldsymbol{a}_l}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}^\top \boldsymbol{x}] - \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}^\top]\mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}]$ with $\mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}] = \sum_{t=1}^n a_{lt}\boldsymbol{x}_t$. Then for any $i, j = 1, 2, ..., n$, we have*

$$\frac{\partial \hat{\boldsymbol{V}}_l}{\partial W_{\boldsymbol{Q},ij}} = x_{li} \boldsymbol{W}_{\boldsymbol{K},j}^\top \, Var_{a_l}(\boldsymbol{x}) \boldsymbol{W}_{\boldsymbol{V}} \propto Var_{a_l}(\boldsymbol{x})$$

and

$$\frac{\partial \hat{\boldsymbol{V}}_l}{\partial W_{\boldsymbol{K},ij}} = x_{li} \boldsymbol{W}_{\boldsymbol{Q},j}^\top \, Var_{a_l}(\boldsymbol{x}) \boldsymbol{W}_{\boldsymbol{V}} \propto Var_{a_l}(\boldsymbol{x}).$$

The same is true for $\hat{\boldsymbol{V}}$ of the fully-connected self-attention.

Proof. Let's first consider the derivative of $\hat{\boldsymbol{V}}_l$ over $W_{\boldsymbol{Q},ij}$. Via some algebraic computation, we have

$$\frac{\partial \hat{\boldsymbol{V}}_l}{\partial W_{\boldsymbol{Q},ij}} = \frac{\partial (\boldsymbol{a}_l \boldsymbol{V})}{\partial W_{\boldsymbol{Q},ij}} = \sum_{t=1}^n a_{lt} \left(\frac{\partial \mathcal{T}_l^{knn}(t)}{\partial W_{\boldsymbol{Q},ij}} - \sum_{k_1=1}^n a_{lk_1} \frac{\partial \mathcal{T}_l^{knn}(k_1)}{\partial W_{\boldsymbol{Q},ij}} \right) \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{V}}, \tag{1}$$

where we denote $\mathcal{T}_l^{knn}(k)$ as follow for shorthand:

$$\mathcal{T}_l^{knn}(k_1) = \begin{cases} \boldsymbol{x}_l \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^\top \boldsymbol{x}_{k_1}^\top, & \text{if patch } k_1 \text{ is selected in row } l \\ -\infty, & \text{otherwise} \end{cases}$$

Let denote set $\mathcal{S} \doteq \{i : \text{patch } i \text{ is selected in row } l\}$ and then we consider the right-hand-side of (1).

$$\begin{aligned} (1) &= \sum_{t\in\mathcal{S}}^n a_{lt} \left(\frac{\partial \left(\boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^\top \boldsymbol{x}_k^\top\right)}{\partial W_{\boldsymbol{Q},ij}} - \sum_{k_1\in\mathcal{S}} a_{lk_1} \frac{\partial \left(\boldsymbol{x}_i \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^\top \boldsymbol{x}_{k_1}^\top\right)}{\partial W_{\boldsymbol{Q},ij}} \right) \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{V}} \\ &= \sum_{t\in\mathcal{S}}^n a_{lt} \left(x_{1i} \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{K},j} - \sum_{k_1\in\mathcal{S}} a_{lk_1} x_{li} \boldsymbol{x}_{k_1} \boldsymbol{W}_{\boldsymbol{K},j} \right) \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{V}} \\ &= \underbrace{\sum_{t\in\mathcal{S}}^n a_{lt} x_{li} \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{K},j} \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{V}}}_{(a)} - \underbrace{\sum_{t\in\mathcal{S}}^n a_{lt} \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{V}}}_{(b)} \cdot \underbrace{\sum_{k_1\in\mathcal{S}} a_{lk_1} x_{li} \boldsymbol{x}_{k_1} \boldsymbol{W}_{\boldsymbol{K},j}}_{(c)} . \end{aligned} \tag{2}$$

Since $\boldsymbol{a}_l$ is the l-th row of the attention matrix, we have $a_{lt} \geq 0$ and $\sum_t a_{lt} = 1$. It is possible to treat terms (a), (b) and (c) as the expectation of some quantities over t replicates with probability a_{lt}. Then (2) can be further simplified as

$$
\begin{aligned}
(2) &= \mathbb{E}_{\boldsymbol{a}_l}[x_{li}\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{K},j} \cdot \boldsymbol{x}\boldsymbol{W}_{\boldsymbol{V}}] - \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{K},j}] \cdot \mathbb{E}_{\boldsymbol{a}_l}[x_{li}\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{V}}] \\
&= x_{li}\left(\mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{W}_{\boldsymbol{K},j}^\top\boldsymbol{x}^\top \cdot \boldsymbol{x}\boldsymbol{W}_{\boldsymbol{V}}] - \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{W}_{\boldsymbol{K},j}^\top\boldsymbol{x}^\top] \cdot \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{V}}]\right) \\
&= x_{li}\boldsymbol{W}_{\boldsymbol{K},j}^\top\left(\mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}^\top\boldsymbol{x}] - \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}^\top] \cdot \mathbb{E}_{\boldsymbol{a}_l}[\boldsymbol{x}_t]\right)\boldsymbol{W}_{\boldsymbol{V}} \\
&= x_{li}\boldsymbol{W}_{\boldsymbol{K},j}^\top \mathrm{Var}_{\boldsymbol{a}_l}(\boldsymbol{x})\boldsymbol{W}_{\boldsymbol{V}}, \qquad (3)
\end{aligned}
$$

where the second equality uses the fact that $\boldsymbol{x}_t\boldsymbol{W}_{\boldsymbol{K},j}$ is a scalar.

Combing (1)-(3), we have

$$
\frac{\partial \hat{\boldsymbol{V}}_l}{\partial W_{\boldsymbol{Q},ij}} = x_{li}\boldsymbol{W}_{\boldsymbol{K},j}^\top \mathrm{Var}_{\boldsymbol{a}_l}(\boldsymbol{x})\boldsymbol{W}_{\boldsymbol{V}} \propto \mathrm{Var}_{\boldsymbol{a}_l}(\boldsymbol{x}). \qquad (4)
$$

Due the symmetric on $\boldsymbol{Q}$ and $\boldsymbol{K}$, we can follow the similar procedure to show

$$
\frac{\partial \hat{\boldsymbol{V}}_l}{\partial W_{\boldsymbol{K},ij}} = x_{li}\boldsymbol{W}_{\boldsymbol{Q},j}^\top \mathrm{Var}_{\boldsymbol{a}_l}(\boldsymbol{x})\boldsymbol{W}_{\boldsymbol{V}} \propto \mathrm{Var}_{\boldsymbol{a}_l}(\boldsymbol{x}). \qquad (5)
$$

Finally, by setting $k = n$, one may verify that equations (4) and (5) also hold for fully-connected self-attention.

Proof for Lemma 2 Before given the formal statement of the Lemma 2, we first show the assumptions.

Assumption 2

1. The token $\boldsymbol{x}_i$ is the sub-gaussian random vector with mean $\boldsymbol{\mu}_i$ and variance $(\sigma^2/d)I$ for $i = 1, 2, ..., n$.
2. $\boldsymbol{\mu}$ follows a discrete distribution with finite values $\boldsymbol{\mu} \in \mathcal{V}$. Moreover, there exist $0 < \nu_1, 0 < \nu_2 < \nu_4$ such that a) $\|\boldsymbol{\mu}_i\| = \nu_1$, and b) $\boldsymbol{\mu}_i\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{W}_{\boldsymbol{K}}^T\boldsymbol{\mu}_i \in [\nu_2, \nu_4]$ for all i and $|\boldsymbol{\mu}_i\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{\mu}_j^\top| \leq \nu_2$ for all $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \in \mathcal{V}$.
3. $\boldsymbol{W}_V$ and $\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{W}_{\boldsymbol{K}}^\top$ are element-wise bounded with ν_5 and ν_6 respectively, that is, $|\boldsymbol{W}_V^{(ij)}| \leq \nu_5$ and $|(\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{W}_{\boldsymbol{K}}^\top)^{(ij)}| \leq \nu_6$, for all i, j from 1 to d.

In Assumption 2 we ensure that for a given query patch, the difference between the clustering center and noises are large enough to be distinguished.

Lemma 2 (formal statement of Lemma 2). *Let patch $\boldsymbol{x}_i$ be σ^2-subgaussian random variable with mean $\boldsymbol{\mu}_i$ and there are k_1 patches out of all k patches follows the same clustering center of query l. Per Assumption 2, when $\sqrt{d} \geq 3(\psi(\delta, d) + \nu_2 + \nu_4)$, then with probability $1 - 5\delta$, we have*

$$
\left\| \frac{\sum_{i=1}^k \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_l\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{W}_k^\top\boldsymbol{x}_i\right)\boldsymbol{x}_i\boldsymbol{W}_V}{\sum_{j=1}^k \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_l\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{x}_j\right)} - \boldsymbol{\mu}_l\boldsymbol{W}_V \right\|_\infty \qquad (6)
$$

$$
\begin{aligned}
&\leq 4\exp\left(\frac{\psi(\delta, d)}{\sqrt{d}}\right)\sigma\nu_5\sqrt{\frac{2}{dk}\log\left(\frac{2d}{\delta}\right)} \\
&+ \left[8\exp\left(\frac{\nu_2 - \nu_4 + \psi(\delta, d)}{\sqrt{d}}\right) - \left(7 + \exp\left(\frac{\nu_2 - \nu_4 + \psi(\delta, d)}{\sqrt{d}}\right)\right)\frac{k_1}{k}\right]\|\boldsymbol{\mu}_1\boldsymbol{W}_V\|_\infty,
\end{aligned}
$$

where $\psi(\delta, d) = 2\sigma\nu_1\nu_6\sqrt{2\log\left(\frac{1}{\delta}\right)} + 2\sigma^2\nu_6\log\left(\frac{d}{\delta}\right)$.

Proof. Without loss of generality, we assume the first k patch are the top-k selected patches. From Assumption **2.1**, we can decompose $\boldsymbol{x}_i = \boldsymbol{\mu}_i + \boldsymbol{h}_i$, $i = 1, 2, ..., k$, where $\boldsymbol{h}_i$ is the sub-gaussian random vector with zero mean. We then analyze the numerator part.

$$\begin{aligned}
&\sum_{i=1}^{k}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_l\boldsymbol{W_Q}\boldsymbol{W}_k^\top\boldsymbol{x}_i\right)\boldsymbol{x}_i\boldsymbol{W}_V \\
=&\overbrace{\sum_{i=1}^{k}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_l\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{\mu}_i^\top\right)\boldsymbol{\mu}_i\boldsymbol{W}_v}^{(a)} + \overbrace{\sum_{i=1}^{k}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_l\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{x}_i^\top\right)\boldsymbol{h}_i\boldsymbol{W}_v}^{(b)} \\
&+\overbrace{\sum_{i=1}^{k}\left[\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_l\boldsymbol{W_Q}\boldsymbol{W}_k^\top\boldsymbol{x}_i\right) - \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_l\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{\mu}_i^\top\right)\right]\boldsymbol{\mu}_i\boldsymbol{W}_v}^{(c)}\,. \qquad (7)
\end{aligned}$$

Below we will bound (a), (b) and (c) separately.

Upper bound for (a)**.** Let denote index set $\mathcal{S}_1 = \{i : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_i,\ i = 1, 2, ..., k\}$. We then have

$$\begin{aligned}
&\left\|(a) - \sum_{i\in\mathcal{S}_1}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{x}_i^\top\right)\boldsymbol{\mu}_1\boldsymbol{W}_V\right\|_\infty \\
\leq&(k - |\mathcal{S}_1|)\max_i\left\{\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{x}_i^\top\right)\right\}\|\boldsymbol{\mu}_1\boldsymbol{W}_V\|_\infty \\
\leq&(k - k_1)\exp\left(\frac{\nu_2}{\sqrt{d}}\right)\|\boldsymbol{\mu}_1\boldsymbol{W}_V\|_\infty, \qquad (8)
\end{aligned}$$

where last inequality is from the Assumption **2.2**.

Upper bound for (b)**.** Since each dimension in $\boldsymbol{h}_l$ is the i.i.d random vector with zero mean variance σ^2/d based on Assumption **2.1**, we can use Hoeffding Inequality to derive the following result holds with probability $1 - \delta$.

$$\left\|\sum_{i=1}^{k}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{x}_i^\top\right)\boldsymbol{h}_i\boldsymbol{W}_V\right\|_\infty \leq \sigma\nu_5\sqrt{\frac{2(k_1U_1^2 + (k - k_1)U_2^2)}{d}\log\left(\frac{2d}{\delta}\right)}, \qquad (9)$$

where

$$U_1 = \max_{i\in\mathcal{S}_1}\left\{\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1\boldsymbol{W_Q}\boldsymbol{W}_{\boldsymbol{K}}^\top\boldsymbol{x}_i^\top\right)\right\}$$

and

$$U_2 = \max_{i \notin \mathcal{S}_1} \left\{ \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top \right) \right\}.$$

We then build the upper bound for U_1 and U_2. Since $\boldsymbol{x}_i = \boldsymbol{\mu}_i + \boldsymbol{h}_i$ for $i = 1, 2, ..., k$, we have

$$\begin{aligned} |\boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top| \leq |\boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top| + |\boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{h}_i^\top| \\ + |\boldsymbol{h}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top| + |\boldsymbol{h}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{h}_i^\top| \end{aligned} \tag{10}$$

Via Assumption **2.3** and Hoeffding Inequality, with probability $1 - 4\delta$, the follow results hold.

$$|\boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{h}_i^\top| \leq \sigma \nu_1 \nu_6 \sqrt{2 \log\left(\frac{1}{\delta}\right)} \tag{11}$$

$$|\boldsymbol{h}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top| \leq \sigma \nu_1 \nu_6 \sqrt{2 \log\left(\frac{1}{\delta}\right)} \tag{12}$$

$$|\boldsymbol{h}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{h}_i^\top| \leq 2\sigma^2 \nu_6 \log\left(\frac{d}{\delta}\right) \tag{13}$$

and we denote $\psi(\delta, d) = 2\sigma\nu_1\nu_6 \sqrt{2\log\left(\frac{1}{\delta}\right)} + 2\sigma^2\nu_6 \log\left(\frac{d}{\delta}\right)$ for shorthand and then we have

$$U_1 \leq \exp\left[\frac{1}{\sqrt{d}}\left(\nu_2 + \psi(\delta, d)\right)\right]$$
$$U_2 \leq \exp\left[\frac{1}{\sqrt{d}}\left(\nu_4 + \psi(\delta, d)\right)\right]$$

As a result, with a probability $1 - 5\delta$, we have:

$$\|(b)\|_\infty \leq \sigma \nu_5 \exp\left(\frac{\psi(\delta, d)}{\sqrt{d}}\right) \sqrt{\frac{2(k_1 \exp\left(\frac{2\nu_2}{\sqrt{d}}\right) + (k - k_1)\exp\left(\frac{2\nu_4}{\sqrt{d}}\right))}{d} \log\left(\frac{2d}{\delta}\right)}. \tag{14}$$

Upper bound for (c).

$$\|(c)\|_\infty \leq \left| \sum_{i=1}^{k} \left[\exp\left(\frac{1}{\sqrt{d}} \boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top\right) - \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top\right) \right] \right| \|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_\infty$$

and it implies

$$\begin{aligned}
&\frac{\|(c)\|_\infty}{\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_\infty} \\
&\leq \left| \sum_{i \notin \mathcal{S}_1} \left[\left(\exp\left(\frac{1}{\sqrt{d}} \left(\boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top - \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top \right) \right) - 1 \right) \right] \right| \exp\left(\frac{\nu_2}{\lambda} \right) \\
&+ \left| \sum_{i \in \mathcal{S}_1} \left[\left(\exp\left(\frac{1}{\sqrt{d}} \left(\boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top - \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top \right) \right) - 1 \right) \right] \right| \exp\left(\frac{\nu_5}{\lambda} \right).
\end{aligned} \tag{15}$$

Combine (15) with (11)-(13) and we have with probability $1 - 4\delta$:

$$\|(c)\|_\infty \leq \left| \exp\left(\frac{\psi(\delta, d)}{\sqrt{d}} \right) - 1 \right| \left[(k - k_1) \exp\left(\frac{\nu_2}{\sqrt{d}} \right) + k_1 \exp\left(\frac{\nu_5}{\sqrt{d}} \right) \right] \|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_\infty. \tag{16}$$

From (7), (8), (14) and (16), with probability $1 - 5\delta$, we have

$$\begin{aligned}
&\left\| \sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top \right) \boldsymbol{x}_i \boldsymbol{W}_V - \sum_{i \in \mathcal{S}_1} \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top \right) \boldsymbol{\mu}_1 \boldsymbol{W}_V \right\|_\infty \\
&\leq \exp\left(\frac{\psi(\delta, d)}{\sqrt{d}} \right) \left[\tau_1(k, k_1) \|W_v \boldsymbol{\mu}_1\|_\infty + \sigma \nu_5 \sqrt{\frac{2\tau_2(k, k_1)}{d} \log\left(\frac{2d}{\delta} \right)} \right],
\end{aligned} \tag{17}$$

where $\tau_1(k, k_1) = (k - k_1) \exp\left(\frac{\nu_2}{\sqrt{d}} \right) + k_1 \exp\left(\frac{\nu_5}{\sqrt{d}} \right)$ and $\tau_2(k, k_1) = k_1 \exp\left(\frac{2\nu_2}{\sqrt{d}} \right) + (k - k_1) \exp\left(\frac{2\nu_4}{\sqrt{d}} \right)$.

Now we consider the upper bound the denominator part.

$$\begin{aligned}
&\left| \sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top \right) - \sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top \right) \right| \\
&\leq \overbrace{\left| \sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}} \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top \right) \left[\exp\left(\frac{1}{\sqrt{d}} \left(\boldsymbol{x}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{x}_i^\top - \boldsymbol{\mu}_1 \boldsymbol{W_Q} \boldsymbol{W_K}^\top \boldsymbol{\mu}_i^\top \right) \right) - 1 \right] \right|}^{(g_1)}.
\end{aligned} \tag{18}$$

Via Assumption **2.2**, (11)-(13) and the definition of $\psi(\delta, d)$, with probability $1-5\delta$ the follow results hold.

$$\begin{aligned}(g_1) &\leq \exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left|\sum_{i=1}^{k}\left[\exp\left(\frac{1}{\sqrt{d}}\left[\boldsymbol{x}_1\boldsymbol{W_Q}\boldsymbol{W_K}^{\top}\boldsymbol{x}_i^{\top} - \boldsymbol{\mu}_1\boldsymbol{W_Q}\boldsymbol{W_K}^{\top}\boldsymbol{\mu}_i^{\top}\right]\right)-1\right]\right| \\ &\leq \exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left|\sum_{i=1}^{k}\left[\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)-1\right]\right| \\ &\leq k\exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left[\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)-1\right]. \end{aligned} \tag{19}$$

Combining (18), (19) and Assumption **2.2**, we have

$$\begin{aligned}&\sum_{i=1}^{k}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1\boldsymbol{W_Q}\boldsymbol{W_K}^{\top}\boldsymbol{x}_i^{\top}\right) \\ \geq &\sum_{i=1}^{k}\exp\left(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_1\boldsymbol{W_Q}\boldsymbol{W_K}^{\top}\boldsymbol{\mu}_i^{\top}\right) - k\exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left(\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)-1\right) \\ \geq & k\exp\left(-\frac{\nu_2}{\sqrt{d}}\right) - k\exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left(\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)-1\right). \end{aligned} \tag{20}$$

When $\sqrt{d} \geq 3(\psi(\delta,d)+\nu_2+\nu_4)$, one may verify

$$\begin{aligned}&\exp\left(\frac{\psi(\delta,d)+\nu_2+\nu_4}{\sqrt{d}}\right) - \exp\left(\frac{\nu_2+\nu_4}{\sqrt{d}}\right) \leq \exp\left(\frac{1}{3}\right) - 1 \approx 0.39 < 1 \\ \Rightarrow &\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right) - 1 - \exp\left(-\frac{\nu_2+\nu_4}{\sqrt{d}}\right) < 0 \\ \Rightarrow &\exp\left(-\frac{\nu_2}{\sqrt{d}}\right) - \exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left(\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)-1\right) > 0 \\ \Rightarrow & k\exp\left(-\frac{\nu_2}{\sqrt{d}}\right) - k\exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left(\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)-1\right) > 0. \end{aligned} \tag{21}$$

Finally, via (17), we have

$$
\begin{aligned}
&\left\|\frac{\sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_i^{\top}\right)}{\sum_{j=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_j^{\top}\right)} \boldsymbol{x}_i \boldsymbol{W}_V - \boldsymbol{\mu}_1 \boldsymbol{W}_V\right\|_{\infty} \cdot \sum_{j=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_j^{\top}\right) \\
&= \left\|\sum_{j=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_i^{\top}\right) \boldsymbol{x}_i \boldsymbol{W}_V - \sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_j^{\top}\right) \boldsymbol{\mu}_1 \boldsymbol{W}_V\right\|_{\infty} \\
&\leq \left\|\sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_i^{\top}\right) \boldsymbol{x}_i \boldsymbol{W}_V - \sum_{i \in \mathcal{S}_1} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{\mu}_i^{\top}\right) \boldsymbol{\mu}_1 \boldsymbol{W}_V\right\|_{\infty} \\
&+ \left\|\sum_{i \notin \mathcal{S}_1} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{\mu}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{\mu}_i^{\top}\right) \boldsymbol{\mu}_1 \boldsymbol{W}_V\right\|_{\infty} \\
&\leq \exp\left(\frac{\psi(\delta, d)}{\lambda}\right)\left[2\tau_1(k, k_1)\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty} + \sigma\nu_5\sqrt{\frac{2\tau_2(k, k_1)}{d}\log\left(\frac{2d}{\delta}\right)}\right] \\
&-k_1 \exp\left(\frac{\nu_4}{\sqrt{d}}\right)\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty} + (k - k_1)\exp\left(\frac{\nu_2}{\sqrt{d}}\right)\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty}. \qquad (22)
\end{aligned}
$$

Per (22) and (21), we have

$$
\begin{aligned}
&\left\|\frac{\sum_{i=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_i^{\top}\right)}{\sum_{j=1}^{k} \exp\left(\frac{1}{\sqrt{d}}\boldsymbol{x}_1 \boldsymbol{W}_{\boldsymbol{Q}} \boldsymbol{W}_{\boldsymbol{K}}^{\top} \boldsymbol{x}_j^{\top}\right)} \boldsymbol{x}_i \boldsymbol{W}_V - \boldsymbol{\mu}_1 \boldsymbol{W}_V\right\|_{\infty} \\
&\leq \left(k \exp\left(-\frac{\nu_2}{\sqrt{d}}\right) - k \exp\left(\frac{\nu_4}{\sqrt{d}}\right)\left(\exp\left(\frac{\psi(\delta, d)}{\sqrt{d}}\right) - 1\right)\right)^{-1} \\
&\cdot \left(\exp\left(\frac{\psi(\delta, d)}{\sqrt{d}}\right)\left[2\tau_1(k, k_1)\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty} + \sigma\nu_5\sqrt{\frac{2\tau_2(k, k_1)}{d}\log\left(\frac{2d}{\delta}\right)}\right]\right. \\
&\qquad \left. +(k - k_1)\exp\left(\frac{\nu_2}{\sqrt{d}}\right)\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty} - k_1 \exp\left(\frac{\nu_4}{\sqrt{d}}\right)\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty}\right) \\
&\leq \frac{\exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)\sigma\nu_5\sqrt{\exp\left(\frac{-2\nu_4}{\sqrt{d}}\right)\frac{2\tau_2(k,k_1)}{dk^2}\log\left(\frac{2d}{\delta}\right)}}{1 + \exp\left(-\frac{\nu_4+\nu_2}{\sqrt{d}}\right) - \exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)} \\
&+ \frac{\exp\left(\frac{\psi(\delta,d)-\nu_4}{\sqrt{d}}\right)\frac{2\tau_1(k,k_1)}{k} - \frac{k_1}{k}\exp\left(\frac{\nu_4}{\sqrt{d}}\right) + (1 - \frac{k_1}{k})\exp\left(\frac{\nu_4}{\sqrt{d}}\right)}{1 + \exp\left(-\frac{\nu_4+\nu_2}{\sqrt{d}}\right) - \exp\left(\frac{\psi(\delta,d)}{\sqrt{d}}\right)} \cdot \|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty} \\
&\leq \left[8\exp\left(\frac{\nu_2 - \nu_4 + \psi(\delta, d)}{\sqrt{d}}\right) - \left(7 + \exp\left(\frac{\nu_2 - \nu_4 + \psi(\delta, d)}{\sqrt{d}}\right)\right)\frac{k_1}{k}\right]\|\boldsymbol{\mu}_1 \boldsymbol{W}_V\|_{\infty} \\
&+ 4\exp\left(\frac{\psi(\delta, d)}{\sqrt{d}}\right)\sigma\nu_5\sqrt{\frac{2}{dk}\log\left(\frac{2d}{\delta}\right)}, \qquad (23)
\end{aligned}
$$

where second inequality uses $\nu_4 \geq \mu_2$ and last inequality uses the definition of $\tau_1(k,k_1)$, $\tau_2(k,k_1)$ and $\sqrt{d} \geq 3(\psi(\delta,d)+\nu_2+\nu_4)$.

Proof of Lemma 3

Without loss of generality, we only consider the case with first query patches. In the k-NN attention scheme, we first use the dot-product product to compute the similarity between query and each key-patches and then use the softmax to normalize the similarities. We make the following assumption to facility our analysis.

Assumption 3.1 There exists $\boldsymbol{\beta}^* \in \mathbb{R}^{1\times n}$ and $\boldsymbol{\beta}^* \in \varDelta$ such that $\boldsymbol{q}_1 = \boldsymbol{\beta}\boldsymbol{K} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is filled with random variable follows $\mathcal{N}(0,\sigma^2)$ for some $\sigma > 0$.

To see the connection between the Assumption **3.1** with attention scheme, we consider the follow problem.

$$\min_{\boldsymbol{\beta}\in\varDelta} \|\boldsymbol{K}^\top\boldsymbol{\beta}^\top - \boldsymbol{q}_1^\top\|_2^2, \tag{24}$$

If $\boldsymbol{K}$ is normalized with zero columns mean and we apply the exponential gradient method on the initial solution $\boldsymbol{\beta}_0 = \frac{1}{n}\boldsymbol{e}$ with step length $1/\sqrt{d}$, the one step updated solution $\boldsymbol{\beta}_1$ is

$$\boldsymbol{\beta}_1 = \frac{\exp(\frac{1}{\sqrt{d}}\boldsymbol{K}\boldsymbol{q}_1^\top)}{\sum_{i=1}^k \exp(\frac{1}{\sqrt{d}}\boldsymbol{k}_i\boldsymbol{q}_1^\top)}. \tag{25}$$

The above equation (25) is just the attention scheme used standard transformer type model. Based on Assumption **3.1**, we can treat $\boldsymbol{\beta}_1$ as an approximation of underlying true parameters $\boldsymbol{\beta}^*$.

On the other hand, it is commonly believed that only part of patches are correlated with the query patch (i.e., with non-zero similarity weights.) and it would be ideal if we could use a computational cheap method to eliminate the irrelevant patches. In this paper, we consider the top-k selection scheme. To see the rationality of the top-k selection, we consider augmenting (24) with L_2 regularization on $\boldsymbol{\beta}$.

$$\begin{aligned}
&\min_{\boldsymbol{\beta}\in\varDelta} \|\boldsymbol{K}^\top\boldsymbol{\beta}^\top - \boldsymbol{q}_1^\top\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2,\\
\Rightarrow &\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{\beta}^\top + \boldsymbol{q}_1^\top) + \lambda\boldsymbol{\beta}^\top + \boldsymbol{\lambda_1} + \boldsymbol{e}\lambda_2 = 0,\\
\Rightarrow &\boldsymbol{\beta}^\top = (\boldsymbol{K}\boldsymbol{K}^\top + \lambda\boldsymbol{I})^{-1}\left(\boldsymbol{K}\boldsymbol{q}_1^\top + \boldsymbol{\lambda_1} + \boldsymbol{e}\lambda_2\right),
\end{aligned}$$

where we use the KKT optimal condition, and $\boldsymbol{\lambda}_1$, λ_2 are Lagrange multipliers to make sure $\boldsymbol{\beta} \in \varDelta$. If λ is large enough and $\boldsymbol{\beta} > 0$, we will have

$$\boldsymbol{\beta}^\top \approx \frac{1}{\lambda}\left(\boldsymbol{K}\boldsymbol{q}_1^\top + \boldsymbol{e}\lambda_2\right) \propto \boldsymbol{K}\boldsymbol{q}_1^\top \tag{26}$$

The above result indicates that we may selection the important elements in $\boldsymbol{\beta}$ (e.t., with large magnitude) by considering rankness in vector $\boldsymbol{K}\boldsymbol{q}_1^\top$.

We then discussion the correctness of the top-k selection with the following regularity assumptions on $\boldsymbol{K}$ and $\boldsymbol{q}_1$.

Assumption 3.2

1. $\boldsymbol{K}$ is normalized with row zero mean. Let $\boldsymbol{\Sigma} = \boldsymbol{K}\boldsymbol{K}^\top$ and $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{K}^\top$. We assume there exist some $c, c_4 > 1$ and $C_1 > 0$ such that the following inequality
$$\mathbb{P}\left(\lambda_{\max}(\tilde{p}^{-1}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top) > c_4 \text{ and } \lambda_{\max}(\tilde{p}^{-1}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top) < \frac{1}{c_4}\right) < \exp(-C_1 d)$$
holds for any $\tilde{p} \times d$ submatrix $\tilde{\boldsymbol{Z}}$ of $\boldsymbol{Z}$ with $cn < \tilde{p} \leq n$.
2. $\mathrm{var}(\boldsymbol{q}_1) = \mathcal{O}(1)$ and for some $\kappa \geq 0$ and $c_5, c_6 > 0$,
$$\min_{i:\beta_i^* > 0} \beta_i \geq \frac{c_5}{d^\kappa} \text{ and } \min_{i:\beta_i^* > 0} \mathrm{cov}(\beta_i^{-1}\boldsymbol{q}_1^\top, \boldsymbol{k}_i^\top) \geq c_6$$
3. There exist some $\tau \in [0, 1)$ and $c_7 > 0$ such that
$$\lambda_{\max}(\boldsymbol{\Sigma}) \leq c_7 d^\tau.$$

Lemma 3 (formal statement of Lemma). *Let's assume only be s keywords are relevant to the query l. Under Assumption* **3.1** *and* **3.2***, when* $2\kappa + \tau < 1$*, with probability* $1 - \mathcal{O}(s\exp(-Cd^{1-2\kappa}/\log d))$*, we have*

$$\sum_{i=1}^{n} \mathbb{1}(\boldsymbol{q}_l\boldsymbol{k}_i^\top \geq \min_{i \in \mathcal{M}^*} \boldsymbol{q}_l\boldsymbol{k}_i^\top) \leq cnd^{2\kappa+\tau-1}, \tag{27}$$

where $\mathcal{M}^* = \{i : \textit{keyword } i \textit{ is relevant to the query } l.\}$ *, and* τ*,* κ*,* c *and* C *are positive constants.*

Proof. Our strategy is the similar to the proof of Theorem 1 in [1].

Based on equation (44) in [1], we have

$$\boldsymbol{K}\boldsymbol{K}^\top = n\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{U}}^\top \mathrm{diag}(\mu_1, ..., \mu_d)\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}, \tag{28}$$

where μ_1,..., μ_d are d eigenvalues of $p^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top$, $\tilde{\boldsymbol{U}} = (I_d, \mathbf{0})_{d\times n}\boldsymbol{U}$, and $\boldsymbol{U}$ is uniformally distributed on the orthogonal group $O(n)$. To facility our further analysis we denote $\boldsymbol{\omega} = \boldsymbol{q}_l\boldsymbol{K}^\top$. By definition of $\boldsymbol{\omega}$ and per Assumption **3.1**, we have

$$\boldsymbol{\omega}^\top = \boldsymbol{K}\boldsymbol{q}_l^\top = \boldsymbol{K}\boldsymbol{K}^\top\boldsymbol{\beta}^\top + \boldsymbol{K}\boldsymbol{\epsilon}^\top \doteq \boldsymbol{\xi} + \boldsymbol{\eta}. \tag{29}$$

We then separately study $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$.

Analysis on $\boldsymbol{\xi}$. We first bound $\boldsymbol{\xi}$ from above. Since $\{\mu_i\}$ is the eigenvalues of $n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top$, we have

$$\mathrm{diag}(\mu_1^2, ..., \mu_d^2) \leq \left[\lambda_{\max}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top)\right]^2 I_d \tag{30}$$

and

$$\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}\tilde{\boldsymbol{U}}^\top \leq \lambda_{\max}(\boldsymbol{\Sigma})I_d$$

There lead to

$$\|\boldsymbol{\xi}\|^2 \leq n^2\lambda_{\max}(\boldsymbol{\Sigma})\left[\lambda_{\max}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top)\right]^2 \cdot \boldsymbol{\beta}^\top\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{U}}^\top\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}. \tag{31}$$

Let Q belongs to the orthogonal group $O(n)$ such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|Q\boldsymbol{e}_1$. Then, it follows from Lemma 1 in [1] that

$$\boldsymbol{\beta}^\top\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{U}}^\top\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|\langle Q^\top\boldsymbol{S}Q\boldsymbol{e}_1, \boldsymbol{e}_1\rangle \overset{(d)}{=} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|\langle\boldsymbol{S}\boldsymbol{e}_1, \boldsymbol{e}_1\rangle, \tag{32}$$

where we use the symbol $\overset{(d)}{=}$ to denote being identical in distribution. By part 3 in Assumption **3.2**, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} \leq \mathrm{var}(\boldsymbol{y}) = \mathcal{O}(1)$, and thus via Lemma 4 in [1], we have for some $C > 0$,

$$\mathbb{P}\left(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{U}}^\top\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} > (d/n)\right) \leq \mathcal{O}\left(\exp(-Cd)\right). \tag{33}$$

Combining with $\lambda_{\max}(\boldsymbol{\Sigma}) = O(d^\tau)$ and $\mathbb{P}(\lambda_{\max}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top) > c_1) \leq \exp\left(-C_1 d\right)$ by parts 1 and 3 in Assumption **3.2** along with union bound, we have

$$\mathbb{P}\left(\|\boldsymbol{\xi}\|^2 \geq \mathcal{O}(d^{1+\tau}n)\right) \leq \mathcal{O}(\exp(-Cd)). \tag{34}$$

We then consider the lower bound on ξ_i for $i \in \mathcal{M}_*$. By (28), we have

$$\xi_i = n\boldsymbol{e}_i^\top\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{U}}^\top\mathrm{diag}(\mu_1, ..., \mu_d)\tilde{\boldsymbol{U}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}. \tag{35}$$

Note that $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_i\| = \sqrt{\mathrm{var}(\boldsymbol{X}_i)} = 1$, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\| = \mathcal{O}(1)$. By part 2 of Assumption **3.2**, there exists some $c > 0$ such that

$$\langle\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_i\rangle = \beta_i\mathrm{cov}(\beta_i^{-1}\boldsymbol{q}_1^\top, \boldsymbol{k}_i) \geq \frac{c}{d^\kappa}. \tag{36}$$

Thus, there exists Q in orthogonal group $O(n)$ such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_i = Q\boldsymbol{e}_1$ and

$$\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} = \langle\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_i\rangle Qe_1 + \mathcal{O}(1)Q\boldsymbol{e}_2. \tag{37}$$

Since $(\mu_1, ..., \mu_d)^\top$ is independent of $\tilde{\boldsymbol{U}}$ by Lemma 1 in [1] and the uniform distribution on the orthogonal group $O(n)$ is invariant under itself, it follows that

$$\xi_i \overset{(d)}{=} \langle\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_i\rangle R_1 + \mathcal{O}(n)R_2 \doteq \xi_{i,1} + \xi_{i,2}, \tag{38}$$

where $\boldsymbol{R} = (R_1, R_2, ..., R_n)^\top = \tilde{\boldsymbol{U}}^\top\mathrm{diag}(\mu_1, ..., \mu_d)\tilde{\boldsymbol{U}}^\top\boldsymbol{e}_1$. We will bound the above two terms $\xi_{i,1}$ and $\xi_{i,2}$ separately. One can verify

$$R_1 \geq \boldsymbol{e}_1^\top\tilde{\boldsymbol{U}}^\top\lambda_{\min}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top)I_d\tilde{\boldsymbol{U}}\boldsymbol{e} = \lambda_{\min}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top)\langle\boldsymbol{S}\boldsymbol{e}_1, \boldsymbol{e}_1\rangle, \tag{39}$$

and thus by part 1 of assumption **3.2**, Lemma 4 in [1], and union bound, we have for some $c, C > 0$,

$$\mathbb{P}(R_1 < cd/n) \leq \mathcal{O}(\exp(-Cd)). \tag{40}$$

Therefore we have for some $c > 0$,

$$\mathbb{P}(\xi_{i,1} \leq cd^{1-\kappa}) \leq \mathcal{O}(\exp(-Cd)) \tag{41}$$

Similarly, we can also show that

$$\mathbb{P}(\|\boldsymbol{R}\|^2 \geq O(d/n)) \leq \mathcal{O}(\exp(-Cd)). \tag{42}$$

Via some analysis, we can show that the distribution of $\tilde{\boldsymbol{R}} = (R_2, ..., R_n)^\top$ is invariant under the orthogonal group $O(n-1)$. Then, it follows that $\tilde{\boldsymbol{R}} \overset{(d)}{=} \|\tilde{\boldsymbol{R}}\|\boldsymbol{W}/\|\boldsymbol{W}\|$, where $\boldsymbol{W} = (W_1, ..., W_{n-1})^\top \sim \mathcal{N}(0, I_{n-1})$, independent of $\|\tilde{\boldsymbol{R}}\|$. Thus, we have

$$R_2 \overset{(d)}{=} \|\tilde{\boldsymbol{R}}\| \frac{W_1}{\|\boldsymbol{W}\|}. \tag{43}$$

And via the Lemma 5 in [1], we can show

$$\mathbb{P}(|\xi_{i,2}| \geq c\sqrt{d}|W|) \leq \mathcal{O}(\exp(-Cd)), \tag{44}$$

where W is $\mathcal{N}(0,1)$-distributed random variable. We then pick $x_d = c\sqrt{2C}d^{1-\kappa}/\sqrt{\log d}$. Then, by standard tail bound, we have

$$\mathbb{P}(c\sqrt{d}|W| \geq x_d) \leq \mathcal{O}(\exp(-Cd^{1-2\kappa}/\log d)), \tag{45}$$

Then

$$\mathbb{P}(|\xi_{i,2}| \geq x_d) \leq \mathcal{O}(\exp(-Cd^{1-2\kappa}/\log d)). \tag{46}$$

It implies that for i with $\beta_i^* > 0$, we have

$$\begin{aligned} &\mathbb{P}\left(\xi_{i,1} - |\xi_{i,2}| \leq cd^{1-\kappa}\right) \leq \mathcal{O}(\exp(-Cd^{1-2\kappa}/\log d)) \\ \Rightarrow &\mathbb{P}\left(\xi_i \leq cd^{1-\kappa}\right) \leq \mathcal{O}(\exp(-Cd^{1-2\kappa}/\log d)). \end{aligned}$$

Next, we examine term $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\top = \boldsymbol{K}\boldsymbol{\epsilon}^\top$. Clearly, we have

$$\boldsymbol{K}^\top\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top \leq \boldsymbol{Z}\lambda_{\max}(\boldsymbol{\Sigma})I_n\boldsymbol{Z}^\top = n\lambda_{\max}(\boldsymbol{\Sigma})\lambda_{\max}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top)I_d. \tag{47}$$

Then, it follows that

$$\|\boldsymbol{\eta}\|^2 = \boldsymbol{\epsilon}\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{\epsilon}^\top \leq n\lambda_{\max}(\boldsymbol{\Sigma})\lambda_{\max}(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^\top)\|\boldsymbol{\epsilon}\|^2. \tag{48}$$

From Assumption **3.1**, we know that $\{\epsilon_i^2/\sigma^2\}$ are i.i.d. χ_1^2-distributed random variables. Thus there exist $c, C > 0$ such that

$$\mathbb{P}(\|\boldsymbol{\epsilon}\|^2 > cd\sigma^2) \leq \exp(-Cd) \tag{49}$$

Along with parts 1 and 3 of Assumption **3.2**, we have

$$\mathbb{P}(\|\boldsymbol{\eta}\|^2 > O(d^{1+\tau}n)) \leq \mathcal{O}(\exp(-Cd)). \tag{50}$$

We then bound $|\eta_i|$ from above. Given that $\boldsymbol{\eta} = K\boldsymbol{\epsilon}^\top \sim \mathcal{N}(0, \sigma^2 KK^\top)$. Hence $\eta_i|_{K=\boldsymbol{K}} \sim \mathcal{N}(0, \mathrm{var}(\eta_i|_{K=\boldsymbol{K}}))$ with $\mathrm{var}(\eta_i|K = \boldsymbol{K}) = \sigma^2 \boldsymbol{e}_i^\top \boldsymbol{K}\boldsymbol{e}_i$.

Let $\mathcal{E}$ be the event $\{\mathrm{var}(\eta_i|\boldsymbol{K}) \leq cd\}$ for some c>0. Then, using the same argument in the previous proof. we can show that

$$\mathbb{P}(\mathcal{E}^c) \leq \mathcal{O}(\exp(-Cd)). \tag{51}$$

Condition on the event $\mathcal{E}$, for all $x > 0$, we have

$$\mathbb{P}(|\eta_i| > x|\boldsymbol{K}) \leq \mathbb{P}(\sqrt{cd}|W| > x) \tag{52}$$

, where W is $\mathcal{N}(0, 1)$ random variable. Via union bound, we have

$$\mathbb{P}(|\eta_i| > x) \leq \mathcal{O}(\exp(-Cd)) + \mathbb{P}(\sqrt{cd}|W| > x) \tag{53}$$

By setting $x = \sqrt{2cC}d^{1-\kappa}/\sqrt{\log d}$, we have

$$\mathbb{P}(\sqrt{cn}|W| > x) \leq \mathcal{O}(\exp(-Cd^{1-2\kappa}/\log d)) \tag{54}$$

Then

$$\mathbb{P}(|\eta_i| > o(d^{1-\kappa})) \leq \mathcal{O}(\exp(-Cd^{1-2\kappa}/\log d)) \tag{55}$$

Finally, we could reach

$$\begin{aligned}
&\mathbb{P}(\min_{i:\beta_i^*>0} \xi_i - |\eta_i| \leq c_1 d^\kappa) \leq \mathcal{O}(s\exp(-Cd^{1-2\kappa}/\log d)) \\
&\Rightarrow \mathbb{P}(\min_{i:\beta_i^*>0} \omega_i \leq c_1 d^\kappa) \leq \mathcal{O}(s\exp(-Cd^{1-2\kappa}/\log d))
\end{aligned}$$

Therefore, with probability $1 - \mathcal{O}(s\exp(-Cd^{1-2\kappa}/\log d))$, the magnitudes of ω_i with $\beta_i^* > 0$ are uniformly at least of order $d^{1-\kappa}$ and for some $c > 0$, we have

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{1}(\omega_k \geq \min_{i:\beta_i^*>0} \omega_i) \leq \frac{cp}{d^{1-2\kappa-\tau}} \leq cnd^{2\kappa+\tau-1} \\
\Rightarrow &\sum_{i=1}^n \mathbb{1}(\boldsymbol{q}_l\boldsymbol{k}_i^\top \geq \min_{i\in\mathcal{M}_*} \boldsymbol{q}_l\boldsymbol{k}_i^\top) \leq cnd^{2\kappa+\tau-1}.
\end{aligned}$$

References

1. Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(5), 849–911 (2008) 12, 13, 14