Supplementary Material of "Designing Robust Transformers using Robust Kernel Density Estimation"

A The Non-parametric Regression Perspective of Self-Attention

Given an input sequence $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D_x}$ of N feature vectors, the self-attention mechanism transforms it into another sequence $\mathbf{H} := [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_N]^\top \in \mathbb{R}^{N \times D_v}$ as follows:

$$\boldsymbol{h}_i = \sum_{j \in [N]} \mathrm{softmax}\Big(\frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\sqrt{D}}\Big)\boldsymbol{v}_j, \text{ for } i = 1, \dots, N. \tag{13}$$

The vectors $\boldsymbol{q}_i$, $\boldsymbol{k}_j$ and $\boldsymbol{v}_j$ are the query, key and value vectors, respectively. They are computed as follows:

$$\begin{aligned}
[\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_N]^\top &:= \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q^\top \in \mathbb{R}^{N \times D}, \\
[\boldsymbol{k}_1, \boldsymbol{k}_2, \dots, \boldsymbol{k}_N]^\top &:= \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K^\top \in \mathbb{R}^{N \times D}, \\
[\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_N]^\top &:= \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V^\top \in \mathbb{R}^{N \times D_v},
\end{aligned} \tag{14}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{D \times D_x}$, $\boldsymbol{W}_V \in \mathbb{R}^{D_v \times D_x}$ are the weight matrices. Equation (13) can be written in the following equivalent matrix form:

$$\mathbf{H} = \mathrm{softmax}\Big(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{D}}\Big)\boldsymbol{V}, \tag{15}$$

where the softmax function is applied to each row of the matrix $(\boldsymbol{Q}\boldsymbol{K}^\top)/\sqrt{D}$. Equation (15) is also called the "softmax attention". Assume we have the key and value vectors $\{\boldsymbol{k}_j, \mathbf{v}_j\}_{j \in [N]}$ that is collected from the data generating process

$$\mathbf{v} = f(\boldsymbol{k}) + \varepsilon, \tag{16}$$

where ε is some noise vectors with $\mathbb{E}[\varepsilon] = 0$, and f is the function that we want to estimate. If $\{\boldsymbol{k}_j\}_{j \in [N]}$ are i.i.d. samples from the distribution $p(\boldsymbol{k})$, and $p(\mathbf{v}, \boldsymbol{k})$ is the joint distribution of $(\mathbf{v}, \boldsymbol{k})$ defined by equation (16), we have

$$f(\boldsymbol{k}) = \mathbb{E}[\mathbf{v}|\boldsymbol{k}] = \int_{\mathbb{R}^D} \mathbf{v} \cdot p(\mathbf{v}|\boldsymbol{k}) d\mathbf{v} = \int_{\mathbb{R}^D} \frac{\mathbf{v} \cdot p(\mathbf{v}, \boldsymbol{k})}{p(\boldsymbol{k})} d\mathbf{v}, \tag{17}$$

We need to obtain estimations for both the joint density function $p(\mathbf{v}, \boldsymbol{k})$ and the marginal density function $p(\boldsymbol{k})$ to obtain function f, one popular approach is the kernel density estimation:

$$\hat{p}_\sigma(\mathbf{v}, \boldsymbol{k}) = \frac{1}{N} \sum_{j \in [N]} k_\sigma\left([\mathbf{v}, \boldsymbol{k}] - [\mathbf{v}_j, \boldsymbol{k}_j]\right) \tag{18}$$

$$\hat{p}_\sigma(\boldsymbol{k}) = \frac{1}{N} \sum_{j \in [N]} k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j), \tag{19}$$

where $[\mathbf{v}, \boldsymbol{k}]$ denotes the concatenation of $\mathbf{v}$ and $\boldsymbol{k}$. k_σ could be isotropic Gaussian kernel: $k_\sigma(\boldsymbol{x} - \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/(2\sigma^2)\right)$, we have

$$\hat{p}_\sigma(\mathbf{v}, \boldsymbol{k}) = \frac{1}{N} \sum_{j \in [N]} k_\sigma(\mathbf{v} - \mathbf{v}_j) k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j). \tag{20}$$

Combining equations (19), (20), and (17), we obtain the NW estimator of the function f as

$$\begin{aligned}
\widehat{f}_\sigma(\boldsymbol{k}) &= \int_{\mathbb{R}^D} \frac{\mathbf{v} \cdot \hat{p}_\sigma(\mathbf{v}, \boldsymbol{k})}{\hat{p}_\sigma(\boldsymbol{k})} d\mathbf{v} \qquad (21) \\
&= \int_{\mathbb{R}^D} \frac{\mathbf{v} \cdot \sum_{j \in [N]} k_\sigma(\mathbf{v} - \mathbf{v}_j) k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j)}{\sum_{j \in [N]} k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j)} d\mathbf{v} \\
&= \frac{\sum_{j \in [N]} k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j) \int \mathbf{v} \cdot k_\sigma(\mathbf{v} - \mathbf{v}_j) d\mathbf{v}}{\sum_{j \in [N]} k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j)} \\
&= \frac{\sum_{j \in [N]} \mathbf{v}_j k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j)}{\sum_{j \in [N]} k_\sigma(\boldsymbol{k} - \boldsymbol{k}_j)}. \qquad (22)
\end{aligned}$$

Now we show how the self-attention mechanism is related to the NW estimator. If the keys $\{\boldsymbol{k}_j\}_{j\in[N]}$ are normalized

$$
\begin{aligned}
\widehat{f}_\sigma(\boldsymbol{q}) &= \frac{\sum_{j\in[N]} \mathbf{v}_j \exp\left(-\|\boldsymbol{q}-\boldsymbol{k}_j\|^2/2\sigma^2\right)}{\sum_{j\in[N]} \exp\left(-\|\boldsymbol{q}-\boldsymbol{k}_j\|^2/2\sigma^2\right)} \\
&= \frac{\sum_{j\in[N]} \mathbf{v}_j \exp\left[-\left(\|\boldsymbol{q}\|^2+\|\boldsymbol{k}_j\|^2\right)/2\sigma^2\right] \exp\left(\boldsymbol{q}^\top \boldsymbol{k}_j/\sigma^2\right)}{\sum_{j\in[N]} \exp\left[-\left(\|\boldsymbol{q}\|^2+\|\boldsymbol{k}_j\|^2\right)/2\sigma^2\right] \exp\left(\boldsymbol{q}^\top \boldsymbol{k}_j/\sigma^2\right)} \\
&= \sum_{j\in[N]} \frac{\exp\left(\boldsymbol{q}^\top \boldsymbol{k}_j/\sigma^2\right)}{\sum_{j\in[N]} \exp(\boldsymbol{q}^\top \boldsymbol{k}_j/\sigma^2)} \mathbf{v}_j \\
&= \sum_{j\in[N]} \operatorname{softmax}\left(\boldsymbol{q}^\top \boldsymbol{k}_j/\sigma^2\right) \mathbf{v}_j.
\end{aligned} \tag{23}
$$

Then estimating the softmax attention is equivalent to estimating $\widehat{f}_\sigma(\boldsymbol{q})$.

B Details on Leveraging Robust KDE on Transformers

For simplicity, we use the Huber loss function as the demonstrating example, which is defined as follows:

$$
\rho(x) := \begin{cases} x^2/2, & 0 \le x \le a \\ ax - a^2/2, & a < x, \end{cases} \tag{24}
$$

where a is a constant. The solution to this robust regression problem has the following form:

Proposition 1. *Assume the robust loss function ρ is non-decreasing in $[0,\infty]$, $\rho(0) = 0$ and $\lim_{x\to 0} \frac{\rho(x)}{x} = 0$. Define $\psi(x) := \frac{\rho'(x)}{x}$ and assume $\psi(0) = \lim_{x\to 0} \frac{\rho'(x)}{x}$ exists and finite. Then the optimal $\hat{p}_{robust}$ can be written as*

$$
\hat{p}_{robust} = \sum_{j\in[N]} \omega_j k_\sigma(\boldsymbol{x}_j, \cdot),
$$

where $\omega = (\omega_1, \cdots, \omega_N) \in \Delta_N$, with each $\omega_j \propto \psi\left(\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}_{robust}\|_{\mathcal{H}_{k_\sigma}}\right)$. Here Δ_n denotes the n-dimensional probability simplex.

Proof. The proof of Proposition 1 is mainly adapted from the proof in Kim & Scott (2012). Here, we provide proof of completeness. For any $p \in \mathcal{H}_{k_\sigma}$, we denote

$$
J(p) = \frac{1}{N} \sum_{j\in[N]} \rho\left(\|k_\sigma(\boldsymbol{x}_j,\cdot) - p\|_{\mathcal{H}_{k_\sigma}}\right).
$$

Then we have the following lemma regarding the Gateaux differential of J and a necessary condition for $\hat{p}_{\text{robust}}$ to be optimal solution of the robust loss objective function in equation (5).

Lemma 1. *Given the assumptions on the robust loss function ρ in Proposition 1, the Gateaux differential of J at $p \in \mathcal{H}_{k_\sigma}$ with incremental $h \in \mathcal{H}_{k_\sigma}$, defined as $\delta J(p; h)$, is*

$$
\delta J(p; h) := \lim_{\tau\to 0} \frac{J(p+\tau h) - J(p)}{\tau} = -\langle V(p), h\rangle_{\mathcal{H}_{k_\sigma}},
$$

where the function $V : \mathcal{H}_{k_\sigma} \to \mathcal{H}_{k_\sigma}$ is defined as:

$$
V(p) = \frac{1}{N} \sum_{j\in[N]} \psi\left(\|k_\sigma(\boldsymbol{x}_j,\cdot) - p\|_{\mathcal{H}_{k_\sigma}}\right) (k_\sigma(\boldsymbol{x}_j,\cdot) - p).
$$

A necessary condition for $\hat{p}_{robust}$ is $V(\hat{p}_{robust}) = 0$.

The proof of Lemma 1 can be found in Lemma 1 of Kim & Scott (2012). Based on the necessary condition for $\hat{p}_{\text{robust}}$ in Lemma 1, i.e., $V(\hat{p}_{\text{robust}}) = 0$, we have

$$\frac{1}{N}\sum_{j\in[N]} \psi\left(\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}_{\text{robust}}\|_{\mathcal{H}_{k_\sigma}}\right)\left(k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}_{\text{robust}}\right) = 0.$$

Direct algebra indicates that $\hat{p}_{\text{robust}} = \sum_{j\in[N]} \omega_j k_\sigma(\boldsymbol{x}_j,\cdot)$ where $\omega = (\omega_1, \cdots, \omega_N) \in \Delta_N$, and $\omega_j \propto \psi\left(\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}_{\text{robust}}\|_{\mathcal{H}_{k_\sigma}}\right)$. As a consequence, we obtain the conclusion of the proposition. $\square$

For the Huber loss function, we have that

$$\psi(x) := \begin{cases} 1, & 0 \le x \le a \\ a/x, & a < x. \end{cases}$$

Hence, when the error $\|k_\sigma(\boldsymbol{x}_j,\cdot), \cdot - \hat{p}_{\text{robust}}\|_{\mathcal{H}_{k_\sigma}}$ is over the threshold a, the final estimator will down-weight the importance of $k_\sigma(\boldsymbol{x}_j,\cdot)$. This is in sharp contrast with the standard KDE method, which will assign uniform weights to all of the $k_\sigma(\boldsymbol{x}_j,\cdot)$. As we mentioned in the main paper, the estimator provided in Proposition 1 is circularly defined, as $\hat{p}_{\text{robust}}$ is defined via ω, and ω depends on $\hat{p}_{\text{robust}}$. Such an issue can be addressed by estimating ω with an iterative algorithm termed as kernelized iteratively re-weighted least-squares (KIRWLS). The algorithm starts with randomly initialized $\omega^{(0)} \in \Delta_n$, and perform the following iterative updates between two steps:

$$\hat{p}^{(k)}_{\text{robust}} = \sum_{j\in[N]} \omega_i^{(k-1)} k_\sigma(\boldsymbol{x}_j,\cdot), \quad \omega_j^{(k)} = \frac{\psi\left(\left\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}^{(k)}_{\text{robust}}\right\|_{\mathcal{H}_{k_\sigma}}\right)}{\sum_{j\in[N]}\psi\left(\left\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}^{(k)}_{\text{robust}}\right\|_{\mathcal{H}_{k_\sigma}}\right)}. \tag{25}$$

Note that, the optimal $\hat{p}_{\text{robust}}$ is the fixed point of this iterative update, and the KIRWLS algorithm converges under standard regularity conditions. Furthermore, one can directly compute the term $\left\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}^{(k)}_{\text{robust}}\right\|_{\mathcal{H}_{k_\sigma}}$ via the reproducing property:

$$\begin{aligned}\left\|k_\sigma(\boldsymbol{x}_j,\cdot) - \hat{p}^{(k)}_{\text{robust}}\right\|^2_{\mathcal{H}_{k_\sigma}} &= -2\sum_{m\in[N]} \omega_m^{(k-1)} k_\sigma(\boldsymbol{x}_m, \boldsymbol{x}_j) + k_\sigma(\boldsymbol{x}_j, \boldsymbol{x}_j) \\ &\quad + \sum_{m\in[N], n\in[N]} \omega_m^{(k-1)}\omega_n^{(k-1)} k_\sigma(\boldsymbol{x}_m, \boldsymbol{x}_n). \end{aligned} \tag{26}$$

Therefore, the weights can be updated without mapping the data to the Hilbert space.

C Fourier Attention with Median of Means

We introduce the Fourier Attention coupled with the Median of Means (MoM) principle and show how this is robust to outliers. For any given function $\phi : \mathbb{R} \to \mathbb{R}$ and radius R, we randomly divide the keys $\{\boldsymbol{k}_i\}_{i\in[N]}$ into B subsets $I_1, \ldots, I_B$ of equal size where $|I_1| = |I_2| = \cdots = |I_B| = \mathcal{S}$. Define $\hat{p}_{R,I_m}(\boldsymbol{q}_l) = \frac{1}{\mathcal{S}}\sum_{i\in I_m}\prod_{j=1}^{D}\phi(\frac{\sin(R(q_{lj}-k_{ij}))}{R(q_{lj}-k_{ij})})$, then the MoM Fourier attention is defined as

$$\hat{\mathbf{h}}_l = \frac{\frac{1}{\mathcal{S}}\sum_{i\in I_m}\mathbf{v}_i\prod_{j=1}^{D}\phi(\frac{\sin(R(q_{lj}-k_{ij}))}{R(q_{ij}-k_{lj})})}{\text{median}\{\hat{p}_{R,I_1}(\boldsymbol{q}_l), \ldots, \hat{p}_{R,I_B}(\boldsymbol{q}_l)\}}, \tag{27}$$

where I_m is the block such that $\hat{p}_R(\boldsymbol{q}_l, \boldsymbol{k})$ achieves its median value. To shed light into the robustness of Transformers that use Eq. (27) as the attention mechanism, we demonstrate that the estimator $\hat{p}_R(\boldsymbol{q}) = \text{median}\{\hat{p}_{R,I_1}(\boldsymbol{q}), \ldots, \hat{p}_{R,I_B}(\boldsymbol{q})\}$ is a robust estimator of the density function $p(\boldsymbol{q})$ of the keys. We first introduce a few notations that are useful for stating this result. Denote $\mathcal{C} = \{1 \le i \le N : k_i \text{ is clean}\}$ and $\mathcal{O} = \{1 \le i \le N : k_i \text{ is outlier}\}$. Then, we have $\mathcal{C}\cap\mathcal{O} = \emptyset$ and $\mathcal{C}\cup\mathcal{O} = \{1, 2, \ldots, N\}$. The following result establishes a high probability upper bound on the sup-norm between $\widehat{p}_R(\boldsymbol{q})$ and $p(\boldsymbol{q})$.

Theorem 1. *Assume that the function ϕ satisfies $\int \phi(\sin(z)/z)z^j dz = 0$ for all $1 \leq j \leq m$ and $\int |\phi(\sin(z)/z)||z|^{m+1}dz < \infty$ for some $m \in \mathbb{N}$. Furthermore, the density function $p(\boldsymbol{q})$ satisfies $\sup_{\boldsymbol{q}} |p(\boldsymbol{q})| < \infty$. The number of blocks B and the number of outliers $|\mathcal{O}|$ are such that $B > (2+\delta)|\mathcal{O}|$ where δ is the failure probability. Then, with $\Delta = \frac{1}{2+\delta} - \frac{|\mathcal{O}|}{B}$ for the radius R sufficiently large and δ sufficiently small, with probability at least $1 - \exp(-2\Delta^2 B)$ we find that*

$$\|\hat{p}_R - p\|_\infty \leq C(\frac{1}{R^{m+1}} + \sqrt{\frac{BR^D \log R \log(2/\delta)}{N}})$$

where C is some universal constant.

Remark 1. The result of Theorem 1 indicates by choosing $R = \mathcal{O}(N^{-\frac{1}{2(m+1)+D}})$, the rate of $\hat{p}_R$ to p under the supremum norm is $\mathcal{O}(N^{-\frac{m+1}{2(m+1)+D}})$. With that choice of R, when N approaches infinity, the MoM estimator $\hat{p}_R$ is a consistent estimator of the clean distribution p of the keys. This confirms the validity of using $\hat{p}_R$ to robustify p and similarly the usage of MoM Fourier attention Eq. (27) as a robust attention for Transformers.

Proof. From the formulation of the MoM estimator $\widehat{p}_R(\boldsymbol{q})$, we obtain the following inequality

$$\{\sup_{\boldsymbol{q}} |\hat{p}_R(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\} \subset \{\sup_{\boldsymbol{q}} \sum_{b=1}^{B} \mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \geq \frac{B}{2}\}$$

This bound indicates that to bound $\mathbb{P}(\|\hat{p}_R(\boldsymbol{q}) - p(\boldsymbol{q})\|_\infty \geq \epsilon)$, it is sufficient to bound $\mathbb{P}(\{\sup_{\boldsymbol{q}} \sum_{b=1}^{B} \mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \geq \frac{B}{2}\})$. Indeed, for each $1 \leq b \leq B$, we find that

$$\mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \leq \mathbf{1}_{\{\sup_{\boldsymbol{q}}\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}}.$$

Therefore, we have

$$\sum_{b=1}^{B} \mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \leq \sum_{b=1}^{B} \mathbf{1}_{\{\sup_{\boldsymbol{q}}\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}},$$

which leads to $\sup_{\boldsymbol{q}} \sum_{b=1}^{B} \mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \leq \sum_{b=1}^{B} \mathbf{1}_{\{\sup_{\boldsymbol{q}}\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}}$. This inequality shows that

$$\mathbb{P}(\{\sup_{\boldsymbol{q}} \sum_{b=1}^{B} \mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \geq \frac{B}{2}\}) \leq \mathbb{P}(\sum_{b=1}^{B} \mathbf{1}_{\{\sup_{\boldsymbol{q}}\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}}).$$

To ease the presentation, we denote $W_b = \mathbf{1}_{\{\sup_{\boldsymbol{q}}\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}}$ and $\mathcal{B} = \{1 \leq b \leq B : I_b \cap \mathcal{O} = \emptyset\}$. Then, the following inequalities hold

$$\begin{aligned}
\sum_{b=1}^{B} \mathbf{1}_{\{\sup_{\boldsymbol{q}}\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} &= \sum_{b \in \mathcal{B}} W_b + \sum_{b \in \mathcal{B}^c} W_b \\
&\leq \sum_{b \in \mathcal{B}} W_b + |\mathcal{O}| \\
&\leq \sum_{b \in \mathcal{B}} (W_b - \mathbb{E}[W_b]) + B \cdot \mathbb{P}(\sup_{\boldsymbol{q}} |\hat{p}_{R,I_1}(\boldsymbol{q}) - p(\boldsymbol{q})| > \epsilon) + |\mathcal{O}|,
\end{aligned}$$

where we assume without loss of generality that $1 \in \mathcal{B}$, which is possible due to the assumption that $B > (2+\delta)|\mathcal{O}|$. By adapting Lemma 1 in Nguyen et al. (2022c) to uniform concentration bound, we have

$$\mathbb{P}(\sup_{\boldsymbol{q}} |\hat{p}_{R,I_1}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq C(\frac{1}{R^{m+1}} + \sqrt{\frac{R^D \log R \log(2/\delta)}{|I_1|}})) \leq \delta.$$

By choose $\epsilon = C(\frac{1}{R^{m+1}} + \sqrt{\frac{R^D \log R \log(2/\delta)}{|I_1|}}))$, then we find that

$$\mathbb{P}(\sup_{\boldsymbol{q}} |\hat{p}_{R,I_1}(\boldsymbol{q}) - p(\boldsymbol{q})| > \epsilon) \leq \frac{\delta}{2(2+\delta)}$$

Collecting the above inequalities leads to

$$\mathbb{P}(\{\sup_{\boldsymbol{q}} \sum_{b=1}^{B} \mathbf{1}_{\{|\hat{p}_{R,I_b}(\boldsymbol{q}) - p(\boldsymbol{q})| \geq \epsilon\}} \geq \frac{B}{2}\}) \leq \exp(-2B\Delta^2),$$

where $\Delta = \frac{1}{2+\delta} - \frac{|\mathcal{O}|}{B}$. As a consequence, we obtain the conclusion of the theorem. □

D Dataset Information

WikiText-103 The dataset[1] contains around $268K$ words and its training set consists of about $28K$ articles with $103M$ tokens, this corresponds to text blocks of about 3600 words. The validation set and test sets consist of 60 articles with $218K$ and $246K$ tokens respectively.

ImageNet We use the full ImageNet dataset that contains $1.28M$ training images and $50K$ validation images. The model learns to predict the class of the input image among 1000 categories. We report the top-1 and top-5 accuracy on all experiments. The following ImageNet variants are test sets that are used to evaluate model performance.

ImageNet-C For robustness on common image corruptions, we use ImageNet-C (Hendrycks & Dietterich, 2019) which consists of 15 types of algorithmically generated corruptions with five levels of severity. ImageNet-C uses the mean corruption error (mCE) as a metric: the smaller mCE means the more robust the model under corruption.

ImageNet-A This dataset contains real-world adversarially filtered images that fool current ImageNet classifiers. A 200-class subset of the original ImageNet-1K's 1000 classes is selected so that errors among these 200 classes would be considered egregious, which cover most broad categories spanned by ImageNet-1K.

ImageNet-O This dataset contains adversarially filtered examples for ImageNet out-of-distribution detectors. The dataset contains samples from ImageNet-22K but not from ImageNet-1K, where samples that are wrongly classified as an ImageNet-1K class with high confidence by a ResNet-50 are selected. We use AUPR (area under precision-recall) as the evaluation metric.

ImageNet-R This dataset contains various artistic renditions of object classes from the original ImageNet dataset, which is discouraged by the original ImageNet. ImageNet-R contains 30,000 image renditions for 200 ImageNet classes, where a subset of the ImageNet-1K classes is chosen.

ImageNet-Sketch This dataset contains 50,000 images, 50 images for each of the 1000 ImageNet classes. The dataset is constructed with Google Image queries "sketch of xxx", where xxx is the standard class name. The search is only performed within the "black and white" color scheme.

E Ablation Studies

In this section, we provide additional results and ablation studies that focus on different design choices for the proposed robust KDE attention mechanisms. The detailed experimental settings can be found in the caption of each table.

[1] www.salesforce.com/products/einstein/ai-research/the-wikitext-dependency-language-modeling-dataset/

Table 5: Perplexity (PPL) and negative likelihood loss (NLL) of our methods (lower part) and baselines (upper part) on WikiText-103 using a medium version of Transformer. The best results are highlighted in bold font and the second best are highlighted in underline. On clean data, Transformer-SPKDE achieves better PPL and NLL than other baselines. Under random swap with outlier words, Transformers with MoM self-attention show much better performance.

Method (median version)	Clean Data		Word Swap	
	Valid PPL/Loss	Test PPL/Loss	Valid PPL/Loss	Test PPL/Loss
Transformer (Vaswani et al., 2017b)	27.90/3.32	29.60/3.37	65.36/4.31	68.12/4.36
Performer (Choromanski et al., 2021)	27.34/3.31	29.51/3.36	64.72/4.30	67.43/4.34
Transformer-MGK (Nguyen et al., 2022b)	27.28/3.31	29.24/3.36	64.46/4.30	67.31/4.33
FourierFormer (Nguyen et al., 2022c)	26.51/3.29	28.01/3.33	63.74/4.28	65.27/4.31
Transformer-RKDE (Huber)	26.12/3.28	27.89/3.32	49.37/3.85	51.22/3.89
Transformer-RKDE (Hampel)	25.87/3.27	27.44/3.31	48.62/3.83	51.03/3.88
Transformer-SPKDE	**25.76/3.27**	**27.35/3.31**	46.91/3.79	49.14/3.84
Transformer-MoM	28.26/3.34	29.98/3.38	45.35/3.75	47.92/3.81
FourierFormer-MoM	27.13/3.31	29.02/3.36	**43.23/3.71**	**44.97/3.74**

Table 6: Test PPL/NLL loss versus the parameter a of Huber loss function defined in Eq. (24) (upper) and Hampel loss function (Kim & Scott, 2012) (lower; we use $2 \times a$ and $3 \times a$ as parameters b and c) on original and word-swapped Wiki-103 dataset. The best results are highlighted in bold font and the second best are highlighted in underline. We choose $a = 0.4$ in rest of the experiments.

Robust Loss Parameter	0.1	0.2	0.4	0.6	0.8	1
Clean Data	32.92/3.48	32.87/3.48	**32.29/3.47**	32.38/3.48	32.46/3.48	32.48/3.48
Word Swap	55.82/3.99	55.97/3.99	**55.68/3.99**	56.89/4.01	57.26/4.01	57.37/4.01
Clean Data	32.67/3.48	**32.32/3.48**	32.35/3.48	32.47/3.48	32.53/3.48	32.58/3.48
Word Swap	58.02/4.03	**57.86/4.03**	57.92/4.03	58.24/4.04	58.37/4.04	58.43/4.04

Table 7: Top-1 classification accuracy on ImageNet versus the parameter a of Huber loss function defined in Eq. (24) under different settings. The best results are highlighted in bold font and the second best are highlighted in underline. We choose $a = 0.2$ in rest of the experiments.

Huber Loss Parameter	0.1	0.2	0.4	0.6	0.8	1
Clean Data	71.45	**72.83**	71.62	71.07	70.65	70.34
FGSM	**56.72**	55.83	55.34	54.87	54.02	52.98
PGD	**46.37**	44.15	43.87	43.25	42.69	41.96
SPSA	52.38	**52.42**	51.69	51.34	50.97	48.22
Imagenet-C	45.37	45.58	**45.63**	45.26	44.63	43.76

Table 8: Top-1 classification accuracy on ImageNet versus the parameter a of Hampel loss function defined in Kim & Scott (2012) under different settings. We use $2 \times a$ and $3 \times a$ as parameters b and c. The best results are highlighted in bold font and the second best are highlighted in underline. We choose $a = 0.2$ in rest of the experiments.

Hampel Loss Parameter	0.1	0.2	0.4	0.6	0.8	1
Clean Data	71.63	**72.94**	71.84	71.23	70.87	70.41
FGSM	**56.42**	55.92	55.83	55.66	54.97	53.68
PGD	**45.18**	44.23	43.89	43.62	43.01	42.34
SPSA	**52.96**	52.48	52.13	51.46	50.92	50.23
Imagenet-C	44.76	45.61	46.04	**46.13**	45.82	45.31

Table 9: Top-1 classification accuracy on ImageNet versus the parameter β of SPKDE defined in Eq. (6) under different settings. $\beta = \frac{1}{1-\varepsilon} > 1$, where ε is the percentage of anomalous samples. A larger β indicates a more robust model. The best results are highlighted in bold font and the second best are highlighted in underline. We choose $\beta = 1.4$ in rest of the experiments.

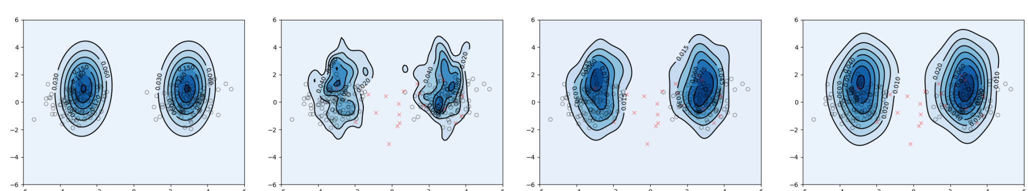
β	1.05	1.2	1.4	1.6	1.8	2
Clean Data	**74.25**	73.56	73.22	73.01	72.86	72.64
FGSM	53.69	55.08	**56.03**	55.37	54.21	53.86
PGD	42.31	43.68	**44.51**	44.32	44.17	43.71
SPSA	51.29	52.02	52.64	**52.84**	52.16	51.39
Imagenet-C	44.68	**45.49**	44.76	44.21	43.96	43.33

Table 10: Top-1 classification accuracy on ImageNet versus the number of iterations of the KIRWLS algorithm in Eq. (25) employed in Transformer-RKDE. Since the increased number of iterations does not lead to significant improvements of performance while the computational cost is much higher, we use the single-step iteration of the KIRWLS algorithm in Transformer-RKDE.

	Huber Loss				Hampel Loss			
Iteration #	1	2	3	5	1	2	3	5
Clean Data	72.83	72.91	72.95	72.98	72.94	72.99	73.01	73.02
FGSM	55.83	55.89	55.92	55.94	55.92	55.96	55.97	55.99
PGD	44.15	44.17	44.17	44.18	44.23	44.26	44.28	44.31
SPSA	52.42	52.44	52.45	52.45	52.48	52.53	52.55	52.56
Imagenet-C	45.58	45.61	45.62	45.62	45.61	45.66	45.68	45.71

Table 11: Computation time (measured by seconds per iteration) of baseline methods, Transformer-SPKDE, Transformer-MoM and Transformer-RKDE with different number of KIRWLS iterations. Transformer-SPKDE requires longer time since it directly obtains the optimal set of weights via the QP solver.

	Iterations of KIRWLS				DeiT	RVT	SPKDE	MoM-KDE
	1	2	3	5				
Time (s/it)	0.43	0.51	0.68	0.84	0.35	0.41	1.45	0.37



(a) (b) (c) (d)

Figure 5: Contour plots of density estimation of the 2-dimensional query vector embedding in an attention layer of the transformer when using (b) KDE (Eq. (4)) and (c) RKDE after one iteration of Eq. (25) with Huber loss (Eq. (24)), (d) KDE with median-of-means principle (Eq. (10)), where (a) is the true density function. We draw 1000 samples (gray circles) from a multivariate normal density and 100 outliers (red cross) from a gamma distribution as the contaminating density. RKDE and KDE with the median-of-means principle can be less affected by contaminated samples when computing self-attention as nonparametric regression.