FourierFormer: Transformer Meets Generalized Fourier Integral Theorem

Tan M. Nguyen∗
Department of Mathematics
University of California, Los Angeles
tanmnguyen89@ucla.edu

Minh Pham∗
Department of Mathematics
University of California, Los Angeles
minhrose@ucla.edu

Tam Nguyen
Department of ECE
Rice University
nguyenminhtam9520@gmail.com

Khai Nguyen
Department of Statistics and Data Sciences
University of Texas at Austin
khainb@utexas.edu

Stanley J. Osher∗∗
Department of Mathematics
University of California, Los Angeles
sjo@math.ucla.edu

Nhat Ho∗∗
Department of Statistics and Data Sciences
University of Texas at Austin
minhnhat@utexas.edu

Abstract

Multi-head attention empowers the recent success of transformers, the state-of-the-art models that have achieved remarkable success in sequence modeling and beyond. These attention mechanisms compute the pairwise dot products between the queries and keys, which results from the use of unnormalized Gaussian kernels with the assumption that the queries follow a mixture of Gaussian distribution. There is no guarantee that this assumption is valid in practice. In response, we first interpret attention in transformers as a nonparametric kernel regression. We then propose the FourierFormer, a new class of transformers in which the dot-product kernels are replaced by the novel generalized Fourier integral kernels. Different from the dot-product kernels, where we need to choose a good covariance matrix to capture the dependency of the features of data, the generalized Fourier integral kernels can automatically capture such dependency and remove the need to tune the covariance matrix. We theoretically prove that our proposed Fourier integral kernels can efficiently approximate any key and query distributions. Compared to the conventional transformers with dot-product attention, FourierFormers attain better accuracy and reduce the redundancy between attention heads. We empirically corroborate the advantages of FourierFormers over the baseline transformers in a variety of practical applications including language modeling and image classification.

1 Introduction

Transformers [83] are powerful neural networks that have achieved tremendous success in many areas of machine learning [40, 76, 36] and become the state-of-the-art model on a wide range of applications across different data modalities, from language [23, 1, 18, 13, 62, 4, 8, 21] to images [24, 43, 78, 63, 59, 27], videos [3, 44], point clouds [97, 31], and protein sequence [65, 34]. In addition to their excellent performance on supervised learning tasks, transformers can also effectively transfer the learned knowledge from a pretraining task to new tasks with limited or no supervision [60, 61, 23, 94, 42]. At the core of transformers is the dot-product self-attention, which

∗ Co-first authors. ∗∗ Co-last authors. Please correspond to: tanmnguyen89@ucla.edu

mainly accounts for the success of transformer models [14, 56, 41]. This dot-product self-attention learns self-alignment between tokens in an input sequence by estimating the relative importance of a given token with respect to all other tokens. It then transforms each token into a weighted average of the feature representations of other tokens where the weight is proportional to a importance score between each pair of tokens. The importance scores in self-attention enable a token to attend to other tokens in the sequence, thus capturing the contextual representation [6, 83, 38].

1.1 Self-Attention

Given an input sequence \( X := [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times D_x} \) of \( N \) feature vectors, self-attention computes the output sequence \( H \) from \( X \) as follows:

**Step 1: Projecting the input sequence into different subspaces.** The input sequence \( X \) is transformed into the query matrix \( Q \), the key matrix \( K \), and the value matrix \( V \) via three linear transformations

\[
Q = X W^Q, \quad K = X W^K, \quad V = X W^V,
\]

where \( W^Q, W^K \in \mathbb{R}^{D \times D_x} \), and \( W^V \in \mathbb{R}^{D \times D_x} \) are the weight matrices. We denote \( Q := [q_1, \ldots, q_N]^\top, K := [k_1, \ldots, k_N]^\top, \) and \( V := [v_1, \ldots, v_N]^\top \), where the vectors \( q_i, k_i, v_i \) for \( i = 1, \ldots, N \) are the query, key, and value vectors, respectively.

**Step 2: Computing the output as a weighted average.** The output sequence \( H := [h_1, \ldots, h_N]^\top \) is then given by

\[
H = \text{softmax}(QK^\top / \sqrt{D})V := AV,
\]

where the softmax function is applied to each row of the matrix \( (QK^\top) / \sqrt{D} \). For each query vector \( q_i, i = 1, \ldots, N \), Eqn. (1) can be written in the vector form to compute the output vector \( h_i \) as follows

\[
h_i = \sum_{j=1}^{N} \text{softmax}(q_i^\top k_j / \sqrt{D})v_j := \sum_{j=1}^{N} a_{ij}v_j.
\]

The matrix \( A \in \mathbb{R}^{N \times N} \) and its component \( a_{ij} \) for \( i, j = 1, \ldots, N \) are the attention matrix and attention scores, respectively. The self-attention computed by equations (1) and (2) is called the dot-product attention or softmax attention. In our paper, we refer a transformer that uses this attention as the baseline transformer with the dot-product attention or softmax attention. In our paper, we refer a transformer that uses this attention as the baseline transformer with the dot-product attention or softmax attention.

**Multi-head Attention** Each output sequence \( H \) forms an attention head. Multi-head attention concatenates multiple heads to compute the final output. Let \( H \) be the number of heads and \( W^O \in \mathbb{R}^{HD \times ND} \) be the projection matrix for the output. The multi-head attention is defined as

\[
\text{MultiHead}(\{Q, K, V\}_{i=1}^{H}) = \text{Concat}(H_1, \ldots, H_H)W^O.
\]

The capacity of the attention mechanism and its ability to learn diverse syntactic and semantic relationships determine the success of transformers [77, 84, 17, 85, 32]. However, equations (1) and (2) implies that the dot-product attention assumes features \( (q_{i1}, \ldots, q_{iD}) \) in \( q_i \), as well as the features \( (k_{j1}, \ldots, q_{jD}) \) in \( k_j \), are independent. Thus, the dot-product attention fails to capture the correlations between these features, limiting its representation capacity and inhibit the performance of transformers on practical tasks where there is no guarantee that independent features can learn from complex data. One solution to capture correlations between features \( q_i \) and \( k_j \) is to introduce covariance matrices into the formulation of the dot-product attention with the cost of significantly increasing of the computational complexity. Also, choosing good covariance matrices is difficult.

1.2 Contribution

In this paper, we first establish a correspondence between self-attention and nonparametric kernel regression. Under this new perspective of self-attention, we explain the limitation of the dot-product self-attention that it may fail to capture correlations between the features in the query and key vectors. We then leverage the generalized Fourier integral theorems, which can automatically capture these correlations, and derive the generalized Fourier integral estimators for the nonparametric regression problem. Using this new density estimator, we propose the FourierFormer, a novel class of transformers that can capture correlations between features in the query and key vectors of self-attention. In summary, our contribution is three-fold:
1. We derive the formula of self-attention from solving a nonparametric kernel regression problem, thus providing a nonparametric regression interpretation to study and further develop self-attention.

2. We develop the generalized Fourier integral estimators for the nonparametric regression problem and provide theoretical guarantees for these estimator.

3. We propose the FourierFormer whose attentions use the generalized Fourier integral estimators to capture more efficiently correlations between features in the query and key vectors.

Finally, we empirically show that the FourierFormer attains significantly better accuracy than the baseline transformer with the dot-product attention on a variety of tasks including the WikiText language modeling and ImageNet image classification. We also demonstrate in our experiments that FourierFormer helps reduce the redundancy between attention heads.

**Organization** We structure this paper as follows: In Section 2, we present the correspondence between self-attention and nonparametric kernel regression. In Section 3, we discuss the generalized Fourier integral estimators and define the FourierFormer. We validate and empirically analyze the advantages of FourierFormer in Section 4. We discuss related works in Section 5. The paper ends with concluding remarks. Technical proofs and more experimental details are provided in the Appendix.

**Notation** For any $N \in \mathbb{N}$, we denote $[N] = \{1, 2, \ldots, N\}$. For any $D \geq 1$, $L_1(\mathbb{R}^D)$ denotes the space of real-valued functions on $\mathbb{R}^D$ that are integrable. For any two sequences $\{a_N\}_{N \geq 1}, \{b_N\}_{N \geq 1}$, we denote $a_N = O(b_N)$ to mean that $a_N \leq C b_N$ for all $N \geq 1$ where $C$ is some universal constant.

## 2 A Nonparametric Regression Interpretation of Self-attention

In this section, we establish the connection between self-attention and nonparametric kernel regression. In particular, we derive the self-attention in equation (2) as a nonparametric kernel regression in which the key vectors $k_i$ and value vectors $v_j$ are training inputs and training targets, respectively, while the query vectors $q_i$ and the output vectors $h_i$ form a set of new inputs and their corresponding targets that need to be estimated, respectively, for $i, j = 1, \cdots, N$. In general, we can view the training set $\{k_j, v_j\}$ for $j \in [N]$ to come from the following *nonparametric regression model*:

$$v_j = f(k_j) + \varepsilon_j,$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are independent noises such that $\mathbb{E}(\varepsilon_j) = 0$. Furthermore, we consider a random design setting where the key vectors $k_1, k_2, \ldots, k_N$ are i.i.d. samples from the distribution that admits $p$ as density function. By an abuse of notation, we also denote $p$ as the joint density where the key and value vectors $(v_1, k_1), \ldots, (v_N, k_N)$ are i.i.d. samples from. Here, $f$ is a true but unknown function and we would like to estimate it.

**Nadaraya–Watson estimator** Our approach to estimate the function $f$ is based on Nadaraya–Watson’s nonparametric kernel regression approach [50]. In particular, from the nonparametric regression model (3), we have $\mathbb{E}[v_j|k_j] = f(k_j)$ for all $j \in [N]$. Therefore, it is sufficient to estimate the conditional distribution of the value vectors given the key vectors. Given the density function $p$ of the key vectors and the joint density $p$ of the key and value vectors, for any pair of vectors $(v, k)$ generate from model (3) we have

$$\mathbb{E}[v|k] = \int_{\mathbb{R}^D} v \cdot p(v|k)dv = \int \frac{v \cdot p(v, k)}{p(k)} dv.$$  

The formulation (4) of the conditional expectation indicates that as long as we can estimate the joint density function $p(v, k)$ and the marginal density function $p(v)$, we are able to obtain an estimation for the conditional expectation and thus for the function $f$. This approach is widely known as Nadaraya–Watson’s nonparametric kernel regression approach.

**Kernel density estimator** To estimate $p(v, k)$ and $p(k)$, we employ the kernel density estimation approach [66, 57]. In particular, by using the isotropic Gaussian kernel with bandwidth $\sigma$, we have the following estimators of $p(v, k)$ and $p(k)$:

$$\hat{p}_\sigma(v, k) = \frac{1}{N} \sum_{j=1}^{N} \varphi_\sigma(v - v_j) \varphi_\sigma(k - k_j), \quad \hat{p}_\sigma(k) = \frac{1}{N} \sum_{j=1}^{N} \varphi_\sigma(k - k_j),$$
When we choose \( \sigma \), where \( \phi \) is the isotropic multivariate Gaussian density function with diagonal covariance matrix \( \sigma^2 I_D \). Given the kernel density estimators (5), we obtain the following estimation of the function \( f \):

\[
\hat{f}_\sigma(k) = \frac{1}{R^D} \int_{R^D} v \cdot \varphi_\sigma(v, k) dv = \frac{1}{R^D} \int_{R^D} v \cdot \sum_{j=1}^N \varphi_\sigma(v - v_j) \varphi_\sigma(k - k_j) dv
\]

\[
= \sum_{j=1}^N \varphi_\sigma(k - k_j) \int_{R^D} v \cdot \varphi_\sigma(v - v_j) dv \frac{v}{\sum_{j=1}^N \varphi_\sigma(k - k_j)} = \sum_{j=1}^N \frac{v_j \varphi_\sigma(k - k_j)}{\sum_{j=1}^N \varphi_\sigma(k - k_j)}.
\]

(6)

Connection between Self-Attention and nonparametric regression By plugging the query vectors \( q_i \) into the function \( \hat{f}_\sigma \) in equation (6), we obtain that

\[
\hat{f}_\sigma(q_i) = \frac{\sum_j^N v_j \exp \left(-\|q_i - k_j\|^2/(2\sigma^2)\right)}{\sum_j^N \exp \left(-\|q_i - k_j\|^2/(2\sigma^2)\right)}
\]

\[
= \frac{\sum_j^N v_j \exp \left[-\left(\|q_i\|^2 + \|k_j\|^2\right)/(2\sigma^2)\right] \exp \left(q_i k_j^\top/\sigma^2\right)}{\sum_j^N \exp \left[-\left(\|q_i\|^2 + \|k_j\|^2\right)/(2\sigma^2)\right] \exp \left(q_i k_j^\top/\sigma^2\right)}.
\]

(7)

If we further assume that the keys \( k_j \) are normalized, which is usually done in practice to stabilize the training of transformers [71], the value of \( \hat{f}_\sigma(q_i) \) in equation (6) then becomes

\[
\hat{f}_\sigma(q_i) = \frac{\sum_j^N v_j \exp \left(q_i k_j^\top/\sigma^2\right)}{\sum_j^N \exp \left(q_i k_j^\top/\sigma^2\right)} = \sum_{j=1}^N \text{softmax} \left(q_i^\top k_j/\sigma^2\right) v_j.
\]

(8)

When we choose \( \sigma^2 = \sqrt{D} \) where \( D \) is the dimension of \( q_i \) and \( k_j \), equation (8) matches equation (2) of self-attention, namely, \( \hat{f}_\sigma(q_i) = h_i \). Thus, we have shown that self-attention performs nonparametric regression using isotropic Gaussian kernels.

Remark 1 The assumption that \( k_j \) is normalized is to recover the pairwise dot-product attention in transformers. In general, this assumption is not necessary. In fact, the isotropic Gaussian kernel in equation (7) is more desirable than the dot-product kernel in equation (8) of the pairwise dot-product attention since the former is Lipschitz while the later is not Lipschitz [37]. The Lipschitz constraint helps improve the robustness of the model [16, 81, 2] and stabilize the model training [48].

Limitation of Self-Attention From our nonparametric regression interpretation, self-attention is derived from the use of isotropic Gaussian kernels for kernel density estimation and nonparametric regression estimation, which may fail to capture the complex correlations between \( D \) features in \( q_i \) and \( k_j \) [88, 33]. Using multivariate Gaussian kernels with dense covariance matrices can help capture such correlations; however, choosing good covariance matrices is challenging and inefficient [87, 73, 11]. In the following section, we discuss the Fourier integral estimator and its use as a kernel for computing self-attention in order to overcome these limitations.

3 FourierFormer: Transformer via Generalized Fourier Integral Theorem

In the following, we introduce generalized integral theorems that are able to capture the complex interactions among the features of the queries and keys. We then apply these theorems to density estimation and nonparametric regression problems. We also establish the convergence rates of these estimators. Given these density estimators, we introduce a novel family of transformers, named FourierFormer, that integrates the generalized Fourier integral theorem into the dot-product attention step of the standard transformer.

3.1 Generalized Fourier Integral Theorems and Their Applications

The Fourier integral theorem is a beautiful result in mathematics [92, 7] and has been recently used in nonparametric mode clustering, deconvolution problem, and generative modeling [33]. It is a combination of Fourier transform and Fourier inverse transform. In particular, for any function
\( p \in L_1(\mathbb{R}^D) \), the Fourier integral theorem is given by

\[
p(k) = \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} \cos(s^\top (k - y)) \, dp(y) \, ds = \frac{1}{\pi^D} \lim_{R \to \infty} \int_{\mathbb{R}^D} \prod_{j=1}^D \frac{\sin(R(k_j - y_j))}{(k_j - y_j)} \, dp(y) \, dy,
\]

(9)

where \( k = (k_1, \ldots, k_D) \), \( y = (y_1, \ldots, y_D) \), \( s = (s_1, \ldots, s_D) \), and \( R \) is the radius. The detailed derivation of Equation (9) is in Appendix B.3. Equation (9) suggests that \( p_R(k) := \frac{1}{\pi^D} \int_{\mathbb{R}^D} \prod_{j=1}^D \frac{\sin(R(y_j - k_j))}{(y_j - k_j)} \, dp(y) \, dy \) can be used as an estimator of the function \( p \).

**Benefits of the Fourier integral over Gaussian kernel** There are two important benefits of the estimator \( p_R \): (i) it can automatically preserve the correlated structure lying within \( p \) even when \( p \) is very complex and high dimensional function. It is in stark contrast to the standard kernel estimator built based on multivariate Gaussian kernel where we need to choose good covariance matrix in the multivariate Gaussian kernel to guarantee such estimator to work well. We note that as the standard soft-max Transformer is constructed based on the multivariate Gaussian kernel, the issue of choosing good covariance matrix in dot-product transformer is inevitable; (ii) The product of sinc kernels in the estimator \( p_R \) does not decay to a point mass when \( R \to \infty \). It is in stark difference from the multivariate Gaussian kernel estimator, which converges to a point mass when the covariance matrix goes to 0. It indicates that \( p_R \) is a non-trivial estimator of the function \( p \). Finally, detailed illustrations of these benefits of the Fourier integral over Gaussian kernel in density estimation and nonparametric regression problems, which we have just shown to have connection to the self-attention in transformer, can be found in Section 8 in [33].

**Generalized Fourier integral estimator** Borrowing the above benefits of Fourier integral estimator \( p_R \), in the paper we would like to consider a generalization of that estimator, named generalized Fourier integral estimator, which is given by:

\[
p_R^\phi(k) := R_D \int_{\mathbb{R}^D} \frac{1}{A_D} \prod_{j=1}^D \phi \left( \frac{\sin(R(y_j - k_j))}{R(y_j - k_j)} \right) \, dp(y) \, dy,
\]

(10)

where \( A := \int_{\mathbb{R}^D} \phi \left( \frac{\sin(z)}{z} \right) \, dz \) and \( \phi : \mathbb{R} \to \mathbb{R} \) is a given function. When \( \phi(k) = k \) for all \( k \in \mathbb{R}^D \), the generalized Fourier integral estimator \( p_R^\phi \) becomes the Fourier integral estimator \( p_R \). Under appropriate conditions on the function \( \phi \) (see Theorem 1 in Section 3.1.1 and Theorem 3 in Appendix C.1), the estimator \( p_R^\phi \) converges to the true function \( p \), namely,

\[
p(k) = \lim_{R \to \infty} p_R^\phi(k) = \lim_{R \to \infty} R_D \int_{\mathbb{R}^D} \frac{1}{A_D} \prod_{j=1}^D \phi \left( \frac{\sin(R(y_j - k_j))}{R(y_j - k_j)} \right) \, dp(y) \, dy.
\]

(11)

We name the above limit as generalized Fourier integral theorem. Furthermore, the estimator \( p_R^\phi \) also inherits similar aforementioned benefits of the Fourier integral estimator \( p_R \). Therefore, we will use the generalized Fourier integral theorem as a building block for constructing density estimators and nonparametric regression estimators, which are crucial to develop the FourierFormer in Section 3.2.

### 3.1.1 Density Estimation via Generalized Fourier Integral Theorems

We first apply the generalized Fourier integral theorem to the density estimation problem. To ease the presentation, we assume that \( k_1, k_2, \ldots, k_N \in \mathbb{R}^D \) are i.i.d. samples from a distribution admitting density function \( p \) where \( D \geq 1 \) is the dimension. Inspired by the generalized Fourier integral theorem, we obtain the following generalized Fourier density estimator \( p_{N,R}^\phi \) of \( p \) as follows:

\[
p_{N,R}^\phi(k) := \frac{R_D}{NA^D} \sum_{i=1}^N \prod_{j=1}^D \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right),
\]

(12)

where \( A = \int_{\mathbb{R}^D} \phi \left( \frac{\sin(z)}{z} \right) \, dz \) and \( k_i = (k_{i1}, \ldots, k_{iD}) \) for all \( i \in [N] \). To quantify the error between the generalized Fourier density estimator \( p_{N,R}^\phi \) and the true density \( p \), we utilize mean integrated
We now have the following result establishing the mean square error (MSE) of

\[
\text{MISE}(\hat{p}_{N,R}^\phi) := \int_{\mathbb{R}^D} (\hat{p}_{N,R}^\phi(k) - p(k))^2 dk.
\]  

(13)

We start with the following bound on the MISE between \( \hat{p}_{n,R}^\phi \) and \( p \).

**Theorem 1** Assume that \( \int_{\mathbb{R}} \phi(\sin(z)/z)z^j dz = 0 \) for all \( j \in \{0, \ldots, m\} \) and \( \int_{\mathbb{R}} |\phi(\sin(z)/z)|z^m dz < \infty \) for some \( m \in \mathbb{N} \). Then, there exist universal constants \( C \) and \( C' \) depending on \( d \) and \( A \) such that

\[
\text{MISE}(\hat{p}_{N,R}^\phi, p) \leq C \frac{R}{(m+1)^D} + C' R D \frac{N}{N}
\]

Proof of Theorem 1 is in Appendix D.1. A few comments are in order. First, by choosing \( R \) to balance the bias and variance in the bound of MISE in Theorem 1, we have the optimal \( R \) as \( R = O(N^{1/(D+m+1)}) \). With that choice of \( R \), the MISE rate of \( \hat{p}_{N,R}^\phi \) is \( O(N^{-m+1}/(D+m+1)) \). Second, when \( \phi(z) = z^l \) for \( l \geq 4 \) and \( z \in \mathbb{R} \), the assumptions in Theorem 1 are satisfied when \( m = 1 \). Under this case, the MISE rate of \( \hat{p}_{N,R}^\phi \) is \( O(N^{-2}/(D+2)) \). However, these assumptions do not satisfy when \( \phi(z) = z^l \) and \( l \in \{1, 2, 3\} \), which is due to the limitation of the current proof technique of Theorem 1 that is based on Taylor expansion of the estimator \( \hat{p}_{N,R}^\phi \).

To address the limitation of the Taylor expansion technique, we utilize the Plancherel theorem in Fourier analysis to establish the MISE rate of \( \hat{p}_{N,R}^\phi \) when \( \phi(z) = z^l \) and \( l \in \{1, 2, 3\} \). The details of the theoretical analyses for such setting are in Appendix C.

### 3.2 FourierFormer: Transformers with Fourier Attention

Motivated by the preservation of the correlated structure of the function from the generalized Fourier integral theorem as well as the theoretical guarantees of density estimators, in this section we adapt the nonparametric regression interpretation of self-attention in Section 2 and propose the generalized Fourier nonparametric regression estimator in Section 3.2.1. We also establish the convergence properties of that estimator. Then, based on generalized Fourier nonparametric regression estimator, we develop the Fourier Attention and its corresponding FourierFormer in Section 3.2.2.

#### 3.2.1 Nonparametric Regression via Generalized Fourier Integral Theorem

We now discuss an application of the generalized Fourier integral theorems to the nonparametric regression setting (3), namely, we assume that \((v_1, k_1), \ldots, (v_N, k_N)\) are i.i.d. samples from the following nonparametric regression model:

\[
v_j = f(k_j) + \varepsilon_j,
\]

where \( \varepsilon_1, \ldots, \varepsilon_N \) are independent noises such that \( \mathbb{E}(\varepsilon_j) = 0 \) and the key vectors \( k_1, k_2, \ldots, k_N \) are i.i.d. samples from \( p \). Given the generalized Fourier density estimator (12), following the argument in Section 2, the Nadaraya–Watson estimator of the function \( f \) based on the generalized Fourier density estimator is given by:

\[
f_{N,R}(k) := \frac{\sum_{i=1}^N \phi(z) \prod_{j=1}^D \frac{\sin(R(k_j-k_{ij}))}{R(k_j-k_{ij})}}{\sum_{i=1}^N \prod_{j=1}^D \frac{\sin(R(k_j-k_{ij}))}{R(k_j-k_{ij})}}.
\]  

(14)

The main difference between the generalized Fourier nonparametric regression estimator \( f_{N,R} \) in equation (14) and the estimator \( \hat{f}_\sigma \) in equation (6) is that the estimator \( f_{N,R} \) utilizes the generalized Fourier density estimator to estimate the conditional distribution of the value vectors given the key vectors instead of the isotropic Gaussian kernel density estimator as in \( \hat{f}_\sigma \). As we highlighted in Section 3, an important benefit of the generalized Fourier density estimator is that it can capture the complex dependencies of the features of the value vectors and the key vectors while the Gaussian kernel needs to have good covariance matrix to do that, which is computationally expensive in practice.

We now have the following result establishing the mean square error (MSE) of \( f_{N,R} \) when \( D_v = 1 \).
Theorem 2 Assume that \( f_R \phi \left( \frac{\sin(z)}{z} \right) z^j dz = 0 \) for all \( 1 \leq j \leq m \) and \( f_R \left| \phi \left( \frac{\sin(z)}{z} \right) \right| z^j dz < \infty \) for any \( m + 1 \leq j \leq 2m + 2 \) for some \( m \in \mathbb{N} \). Then, for any \( k \in \mathbb{R}^D \), when \( D_n = 1 \) there exist universal constants \( C_1, C_2, C_3, C_4 \) such that the following holds:

\[
\mathbb{E} \left[ (f_{N,R}(k) - f(k))^2 \right] \leq \left( \frac{C_1}{R^{2(m+1)}} + \frac{(f(k) + C_2)R^D}{N} \right) \left( p^2(k) J(R) \right),
\]

where \( J(R) = 1 - \frac{1}{p(z)} \left( \frac{C_3}{R^{2(m+1)}} + \frac{C_4R^2 \log(N,R)}{N} \right) \). Here, the outer expectation is taken with respect to the key vectors \( k_1, \ldots, k_N \) and the noises \( \varepsilon_1, \ldots, \varepsilon_N \).

Proof of Theorem 2 is in Appendix D.3. A few comments with Theorem 2 are in order. First, by choosing \( R \) to balance the bias and variance in the bound of the MSE of the nonparametric generalized Fourier estimator \( f_{N,R} \), we have the optimal radius \( R \) as \( R = O\left(N^{-\frac{1}{2(m+1)+D}}\right) \). With that choice of the optimal radius \( R \), the rate of \( f_{N,R} \) is \( O\left(N^{-\frac{4}{2(m+1)+D}}\right) \). Second, when \( \phi(z) = z^l \) for \( l \geq 6 \), the assumption on the function \( \phi \) of Theorem 2 is satisfied with \( m = 1 \). Under this case, the rate of \( f_{N,R} \) becomes \( O\left(N^{-\frac{4}{3}}\right) \). In Appendix C, we also provide the rate of \( f_{N,R} \) when \( \phi(z) = z^l \) for some \( l \leq 5 \), which includes the original Fourier integral theorem.

3.2.2 FourierFormer

Given the generalized Fourier nonparametric regression estimator \( f_{N,R} \) in equation (14), by plugging the query values \( q_1, \ldots, q_N \) into that function, we obtain the following definition of the Fourier attention:

Definition 1 (Fourier Attention) A Fourier attention is a multi-head attention that does nonparametric regression using the generalized Fourier nonparametric regression estimator \( f_{N,R} \). The output \( h_i \) of the Fourier attention is then computed as

\[
\hat{h}_i := f_{N,R}(q_i) = \frac{\sum_{i=1}^{N} \prod_{j=1}^{D} \phi \left( \frac{\sin(R(q_{id} - k_{jd}))}{R(q_{id} - k_{jd})} \right)}{\sum_{i=1}^{N} \prod_{j=1}^{D} \phi \left( \frac{\sin(R(q_{id} - k_{jd}))}{R(q_{id} - k_{jd})} \right)} \quad \forall \ i \in [N].
\]

Given the Fourier Attention in Definition 1, we then give the definition of FourierFormer as follows.

Definition 2 (FourierFormer) A FourierFormer is a transformer that uses Fourier attention to capture dependency between tokens in the input sequence and the correlation between features in each token.

Remark 2 (The Nonnegativity of the Fourier Kernel) The density estimation via generalized Fourier integral theorem in Section 3.1.1 does not require the generalized Fourier density estimator to be nonnegative. However, empirically, we observe that negative density estimator can cause instability in training the FourierFormer. Thus, in FourierFormer, we choose the function \( \phi \) to be a nonnegative function to enforce the density estimator to be nonnegative. In particular, we choose \( \phi \) to be power functions of the form \( \phi(x) = x^{2m} \), where \( m \) is an positive integer. Note that when \( m = 2 \) and \( m = 4 \), the kernels in our generalized Fourier integral estimators are the well-known Fejer-de la Vallée Poussin and Jackson-de la Vallée Poussin kernels [20].

3.3 An Efficient Implementation of the Fourier Attention

The Fourier kernel is implemented efficiently in the C++/CUDA extension developed by Pytorch [58]. The idea is similar to the function \texttt{cdist} [58], which computes the p-norm distance between each pair of the two collections of row vectors. In our case, we aim to compute kernel functions that represent a Fourier attention in Definition 1. The core of this implementation is the following Fourier metric function \( d_f \):

\[
d_f(q_i, k_j) = \prod_{d=1}^{D} \phi \left( \frac{\sin(R(q_{id} - k_{jd}))}{R(q_{id} - k_{jd})} \right).
\]

We directly implement \( d_f \) as a torch.autograd.Function [58] in which we provide an efficient way to compute forward and backward function (\( d_f \) and gradient of \( d_f \)). While the implementation
Table 1. Perplexity (PPL) on WikiText-103 of FourierFormers compared to the baselines. FourierFormers achieve much better PPL than the baselines.

<table>
<thead>
<tr>
<th>Method</th>
<th>Valid PPL</th>
<th>Test PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline dot-product (small)</td>
<td>33.15</td>
<td>34.29</td>
</tr>
<tr>
<td>FourierFormer (small)</td>
<td>31.86</td>
<td>32.85</td>
</tr>
<tr>
<td>Baseline dot-product (medium)</td>
<td>27.90</td>
<td>29.60</td>
</tr>
<tr>
<td>FourierFormer (medium)</td>
<td>26.51</td>
<td>28.01</td>
</tr>
</tbody>
</table>

of the forward function is straightforward, the backward function is more tricky since we need to optimize the code to compute the gradient of $d_f$ w.r.t to variables $q, k,$ and $R$ all at once. We can develop the backward function with highly parallel computation by exploiting GPU architecture and utilizing the reduction technique. The computational time is comparable to function $c_{distr}$; thus, our FourierFormer implementation is as computationally time-efficient.

4 Experimental Results

In this section, we numerically justify the advantage of FourierFormer over the baseline dot-product transformer on two large-scale tasks: language modeling on WikiText-103 [46] (Section 4.1) and image classification on ImageNet [22, 67] (Section 4.2), time series classification on the UEA benchmark [5] (Section 4.3), and reinforcement learning on the D4RL Benchmark [29] (Section 4.4), and the machine translation on the IWSLT’ 14 De-En [10] (Section 4.5). We aim to show that: (i) FourierFormer achieves better accuracy than the baseline transformer on a variety of practical tasks with different data modalities, and (ii) FourierFormer helps reduce head redundancy compared to the baseline transformer (Section 4.6).

Throughout the section, we compare FourierFormers with the baseline dot-product transformers of the same configuration. In all experiments, we made the constant $R$ in Fourier attention (see equation (16)) to be a learnable scalar and set choose the function $\phi(x) = x^4$ (see Remark 2). All of our results are averaged over 5 runs with different seeds. The details on the models and training are provided in Appendix A. Moreover, additional experiments results are provided in Appendix E.

4.1 Language Modeling on WikiText-103

We report the validation and test perplexity (PPL) of FourierFormer versus the baseline transformer with the dot-product attention in Table 1. FourierFormers attain much better PPL than the baselines in both small and medium configurations. For the small configuration, the improvements of FourierFormer over the baseline are 1.29 PPL in validation and 1.44 PPL in test. For the medium configuration, these improvements are 1.39 PPL in validation and 1.59 PPL in test. These results suggest that the advantage of FourierFormer over the baseline dot-product transformer grows with the model’s size. This meets our expectation because larger models have larger query and key dimensions, e.g. the language model with medium configuration in this experiment has the query and key dimension of 256 versus 128 as in the language model with small configuration. Since the advantage of FourierFormer results from the property that FourierFormer can capture correlation between features in query and key vectors, the larger the query and key dimensions are, the more advantage FourierFormer has.

4.2 Image Classification on ImageNet

In the Imagenet classification task, we illustrate the benefits of Fourierformers in different data modalities. We summarize our models’ results in Table 2. Same as in the language modeling experiment, for this image classification experiment, the Deit model equipped with FourierFormer significantly outperforms the baseline Deit dot-product transformer [79] in both top-1 and top-5 accuracy. This result suggests that the advantage of FourierFormer over the baseline dot-product transformer holds across different data modalities.

4.3 UEA Time Series Classification

To evaluate Fourierformers on temporal sequences, we compare the accuracy of our models and the baseline softmax transformers trained on 10 datasets in the the UEA Time Series Classification Archive benchmark [5]. We summarize our results in Table 3. We observe show that Fourierformers outperforms softmax baselines in 7 out of 10 tasks and yields significantly better accuracy than the softmax transformer on average, showing the our models benefits when trained on temporal data.
Table 2. Top-1 and top-5 accuracy (%) of FourierFormer DeiT vs. the baseline DeiT with dot-product attention. FourierFormer DeiT outperforms the baseline in both top-1 and top-5 accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-1 Acc</th>
<th>Top-5 Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline DeiT</td>
<td>72.23</td>
<td>91.13</td>
</tr>
<tr>
<td>FourierFormer DeiT</td>
<td>73.25</td>
<td>91.66</td>
</tr>
</tbody>
</table>

Table 3. The FourierFormer vs. the baseline softmax transformer on the UEA Time Series Classification Archive benchmark [5]. The FourierFormer outperforms the baseline. We also include the reported results from [95] and [93] (in parentheses) in addition to our reproduced results. The experiment setups and configurations for the baseline and our FourierFormer are the same as in [93] (for the PEMS-SF, SelfRegulationSCP2, UWaveGestureLibrary datasets) and [95] (for other tasks).

<table>
<thead>
<tr>
<th>Dataset/Model</th>
<th>Baseline softmax</th>
<th>FourierFormer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETHANOL CONCENTRATION</td>
<td>32.08 (33.70)</td>
<td>36.12</td>
</tr>
<tr>
<td>FACE DETECTION</td>
<td>68.70 (68.10)</td>
<td>68.71</td>
</tr>
<tr>
<td>HANDWRITING</td>
<td>32.08 (30.50)</td>
<td>31.68</td>
</tr>
<tr>
<td>HEARTBEAT</td>
<td>75.77 (77.60)</td>
<td>76.42</td>
</tr>
<tr>
<td>JAPANESE VOWELS</td>
<td>99.46 (99.40)</td>
<td>99.37</td>
</tr>
<tr>
<td>PEMS-SF</td>
<td>82.66 (82.10)</td>
<td>86.70</td>
</tr>
<tr>
<td>SELFREGULATION SCP1</td>
<td>91.46 (92.50)</td>
<td>91.70</td>
</tr>
<tr>
<td>SELFREGULATION SCP2</td>
<td>54.72 (53.90)</td>
<td>55.37</td>
</tr>
<tr>
<td>SPOKEN ARABIC DIGITS</td>
<td>99.33 (99.30)</td>
<td>99.00</td>
</tr>
<tr>
<td>UWAVE GESTURE LIBRARY</td>
<td>84.45 (85.60)</td>
<td>86.66</td>
</tr>
<tr>
<td><strong>AVERAGE ACCURACY</strong></td>
<td><strong>72.07 (72.27)</strong></td>
<td><strong>73.17</strong></td>
</tr>
</tbody>
</table>

Table 4. The decision FourierFormer vs. the baseline decision transformer [12] on the continuous control tasks from D4RL benchmark [29]. The decision FourierFormer yields significantly better results than the baseline decision transformer on 8 out of 9 tasks and on average across tasks. Each experiment result is averaged over 5 runs with different random seeds. We also include the reported results from [93] (in parentheses) in addition to our reproduced results.

<table>
<thead>
<tr>
<th>Environment/Model</th>
<th>Baseline decision transformer</th>
<th>Decision FourierFormer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDIUM-EXPERT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HALF CHEETAH</td>
<td>91.03 (83.80)</td>
<td>92.27</td>
</tr>
<tr>
<td>HOPPER</td>
<td>110.30 (104.40)</td>
<td>111.10</td>
</tr>
<tr>
<td>WALKER</td>
<td>108.70 (107.70)</td>
<td>108.90</td>
</tr>
<tr>
<td>MEDIUM-REPLAY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HALF CHEETAH</td>
<td>35.31 (34.6)</td>
<td>38.47</td>
</tr>
<tr>
<td>HOPPER</td>
<td>85.61 (79.70)</td>
<td>89.70</td>
</tr>
<tr>
<td>WALKER</td>
<td>66.11 (62.90)</td>
<td>63.19</td>
</tr>
<tr>
<td>MEDIUM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HALF CHEETAH</td>
<td>42.28 (42.40)</td>
<td>42.38</td>
</tr>
<tr>
<td>HOPPER</td>
<td>61.47 (64.20)</td>
<td>64.77</td>
</tr>
<tr>
<td>WALKER</td>
<td>68.68 (70.60)</td>
<td>70.42</td>
</tr>
<tr>
<td><strong>AVG REWARD</strong></td>
<td><strong>74.39 (72.20)</strong></td>
<td><strong>75.69</strong></td>
</tr>
</tbody>
</table>

4.4 Reinforcement learning on the D4RL benchmark

We also examine the performance of our Fourierformers in reinforcement learning. In Table 4, we verify the advantage of decision FourierFormer over the baseline decision transformer [12] on the continuous control tasks from the D4RL benchmark [29]. The decision FourierFormer is the decision transformer with the Fourier attention instead of the softmax attention. On this benchmark, our decision FourierFormer significantly outperforms the baseline decision transformer on 8 out of 9 tasks and on average across tasks. Each experiment result averaged over 5 runs with different random seeds. We follow the architecture and training configuration from [93].

4.5 Machine Translation on IWSLT’14 De-En

We demonstrate the performance of Fourierformer on the IWSLT’14 De-En [10] neural machine translation task, which has different inputs’ the sequence lengths. Table 5 shows that the FourierFormer achieves better BLUE scores than the softmax baseline.

4.6 FourierFormer Helps Reducing Head Redundancy

To study the diversity between attention heads, given the model trained for the WikiText-103 language modeling task, we compute the average $L_2$ distance between heads in each layer. We show the layer-average mean and variance of distances between heads in Table 6. Results in Table 6 shows
Table 5. The FourierFormer vs. the baseline softmax transformer on the IWSLT'14 De-En machine translation benchmark [10]. The FourierFormer outperforms the baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline softmax</td>
<td>34.42</td>
</tr>
<tr>
<td>FourierFormer</td>
<td>34.68</td>
</tr>
</tbody>
</table>

Table 6. Layer-average mean and standard deviation of $L_2$ distances between heads of FourierFormer versus the baseline transformer with dot-product attention trained for the WikiText-103 language modeling task. FourierFormer has greater $L_2$ distance between heads than the baseline and thus captures more diverse attention patterns.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline dot-product</td>
<td>6.20 ± 2.30</td>
<td>6.17 ± 2.30</td>
</tr>
<tr>
<td>FourierFormer</td>
<td>7.45 ± 2.50</td>
<td>7.37 ± 2.44</td>
</tr>
</tbody>
</table>

that FourierFormer obtains greater $L_2$ distance between attention heads than the baseline transformer with the dot-product attention and thus helps reduce the head redundancy. Note that we use the small configuration as specified in Section 4.1 for both models.

5 Related Work

Interpretation of Attention Mechanism in Transformers Recent works have tried to gain an understanding of transformer’s attention from different perspectives. [80] considers attention as applying kernel smoother over the inputs. Extending this kernel approach, [35, 15, 52, 89, 54] linearize the softmax kernel in dot-product attention and propose a family of efficient transformers with linear computational and memory complexity. [9] then shows that these linear transformers are comparable to a Petrov-Galerkin projection [64], suggesting that the softmax normalization in the dot-product attention is sufficient but not necessary. Other works provide an understanding of attention in transformers via ordinary/partial differential equation include [45, 69]. In addition, [51, 75, 30, 96, 53] relate attentions in transformers to a Gaussian mixture models. Several works also connect the attention mechanism to graph-structured learning and message passing in graphical models [90, 72, 39]. Our work focuses on deriving the connection between self-attention and nonparametric kernel regression and exploring better regression estimator, such as the generalized Fourier nonparametric regression estimator, to improve the performance of transformers.

Redundancy in Transformers [19, 47, 25] show that neurons and attention heads in the pre-trained transformer are redundant and can be removed when applied on a downstream task. By studying the contextualized embeddings in pre-trained networks, it has been demonstrated that the learned representations from these redundant models are highly anisotropic [49, 26]. Furthermore, [70, 74, 86, 68] employ knowledge distillation and sparse approximation to enhance the efficiency of transformers. Our FourierFormer is complementary to these methods and can be combined with them.

6 Concluding Remarks

In this paper, we establish the correspondence between the nonparametric kernel regression and the self-attention in transformer. We then develop the generalized Fourier integral estimators and propose the FourierFormer, a novel class of transformers that use the generalized Fourier integral estimators to construct their attentions for efficiently capturing the correlations between features in the query and key vectors. We theoretically prove the approximation guarantees of the generalized Fourier integral estimators and empirically validate the advantage of FourierFormer over the baseline transformer with the dot-product attention in terms of accuracy and head redundancy reduction. It is interesting to incorporate robust kernels into the nonparametric regression framework of FourierFormer to enhance the robustness of the model under data perturbation and adversarial attacks. A limitation of FourierFormer is that it still has the same quadratic computational and memory complexity as the baseline transformer with the dot-product attention. We leave the development of the linear version of FourierFormer that achieves linear computational and memory complexity as future work. It is worth noting that there is no potential negative societal impacts of FourierFormer.

Acknowledgements

This material is based on research sponsored by the AFOSR MURI FA9550-18-1-0502, the ONR grant N00014-20-1-2093, the MURI N00014-20-1-2787, and the NSF under Grant# 2030859 to the Computing Research Association for the CIFellows Project (CIF2020-UCLA-38). NH acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning.
References


Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default [TODO] to [Yes], [No], or [N/A]. You are strongly encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

• Did you include the license to the code and datasets? [Yes] See Section ??.
• Did you include the license to the code and datasets? [No] The code and the data are proprietary.
• Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
Supplement to "FourierFormer: Transformer Meets Generalized Fourier Integral Theorem"

In the supplementary material, we collect proofs, additional theories, and experiment results deferred from the main text. In Appendix C, we provide additional theoretical results for generalized Fourier density estimator and for generalized Fourier nonparametric regression estimator. We provide proofs of key results in the main text and additional theories in Appendix D. We present experiment details in Appendix A while including additional experimental results in Appendix E.

A Experiment Details

This section provides the details of the model and training for experiments in Section 4. All of our experiments are conducted on a server with 4 NVIDIA A100 GPUs.

A.1 Language Modeling

Datasets and metrics WikiText-103 is a collection of articles from Wikipedia, which have long contextual dependencies. The training set consists of about 28K articles containing 103M running words; this corresponds to text blocks of about 3600 words. The validation and test sets have 218K and 246K running words, respectively. Each of them contains 60 articles and about 268K words. Our experiment follows the standard setting [46, 71] and splits the training data into L-word independent long segments. For evaluation, we use a batch size of 1, and process the text sequence with a sliding window of size L. The last position is used for computing perplexity (PPL) except in the first segment, where all positions are evaluated as in [1, 71].

Models and baselines Our implementation is based on the public code by [71].4 We use their small and medium models in our experiments. In particular, for small models, the key, value, and query dimension are set to 128, and the training and evaluation context length are set to 256. For medium models, the key, value, and query dimension are set to 256, and the training and evaluation context length are set to 384. In both configurations, the number of heads is 8, the feed-forward layer dimension is 2048, and the number of layers is 16.

In our experiments on WikiText-103 in Section 4.1, we let R be a learnable scalar initialized to 2 and choose \( \phi(x) = x^4 \). The same setting is used for all attention units in the model; each unit has a different R. We observe that by setting R to be a learnable vector \([R_1, \ldots, R_D]^\top\), the FourierFormer gains advantage in accuracy but with the cost of the increase in the number of parameters. When R is a vector \([R_1, \ldots, R_D]^\top\), the equation of the Fourier Attention is given by

\[
\hat{h}_i := f_{N,R}(q_i) = \frac{\sum_{i=1}^N \prod_{j=1}^D \phi \left( \frac{\sin(R_j(q_{ij} - k_{ij}))}{R_j(q_{ij} - k_{ij})} \right)}{\sum_{i=1}^N \prod_{j=1}^D \phi \left( \frac{\sin(R_j(q_{ij} - k_{ij}))}{R_j(q_{ij} - k_{ij})} \right)} \quad \forall i \in [N]. \tag{16}
\]

We provide an ablation study for the effect of R and \( \phi \) in Section E.

A.2 Image Classification

Datasets and metrics The ImageNet dataset [22, 67] consists of 1.28M training images and 50K validation images. For this benchmark, the model learns to predict the category of the input image among 1000 categories. Top-1 and top-5 classification accuracies are reported.

Models and baselines We use the DeiT-tiny model [79] with 12 transformer layers, 4 attention heads per layer, and a model dimension of 192. To train the models, we follow the same setting and configuration as for the baseline [79].2 Similar to the setting for language modeling, in our experiments on ImageNet image classification, we set R to be a learnable vector \([R_1, \ldots, R_D]^\top\), the FourierFormer gains advantage in accuracy but with the cost of the increase in the number of parameters. When R is a vector \([R_1, \ldots, R_D]^\top\), the equation of the Fourier Attention is given by

\[
\hat{h}_i := f_{N,R}(q_i) = \frac{\sum_{i=1}^N \prod_{j=1}^D \phi \left( \frac{\sin(R_j(q_{ij} - k_{ij}))}{R_j(q_{ij} - k_{ij})} \right)}{\sum_{i=1}^N \prod_{j=1}^D \phi \left( \frac{\sin(R_j(q_{ij} - k_{ij}))}{R_j(q_{ij} - k_{ij})} \right)} \quad \forall i \in [N]. \tag{16}
\]

We provide an ablation study for the effect of R and \( \phi \) in Section E.

A.3 UEA Time Series Classification

Following [93], we choose 10 out of 30 datasets in the benchmark [5], which vary in input sequence lengths, the number of classes, and dimensionality, to evaluate our models on temporal sequences.

---

1Implementation available at https://github.com/IDSIA/lmtool-fwp.
The test accuracy is reported as an evaluation for the benchmark.

**Models and baseline** For all experiments in this task, we adapt the setups and configurations as in [93] (for the PEMS-SF, SelfRegulationSCP2, UWaveGestureLibrary datasets) and [95] (for the other tasks). The number of heads is 8 in all models, whereas the model dimension and number of transformer layers are varied.

### A.4 Reinforcement learning on the D4RL benchmark

**Datasets and metrics** In the D4RL benchmark [29], which consists of the continuous control tasks for offline reinforcement learning, we choose HalfCheetah, Hopper, and Walker as experiment environments and Medium-Expert, Medium, and Medium-Replay as behavior policies. This selection is adapted from [93].

**Models and baseline** The models trained on this benchmark has the same configuration as in [93], with 3 transformer layers and 4 heads per layer. In our D4RL experiments, we choose $\phi = x^4$ and the initial value of the learnable scalar $R$ to be 1.

### A.5 Machine Translation

**Datasets and metrics** The IWSLT'14 De-En dataset consists of $170K$ training sentence pairs, $7K$ validation pairs, and $7K$ test pairs. In this task, the model does the translation from German to English. To measure the performance of the trained model, the BLEU score [55] is used.

**Models and baselines** The architecture of the Fourierformer and the baseline contains 12 transformer layers with 4 heads per layer. Our implementation is based on the public code [link](https://github.com/pytorch/fairseq/tree/main/examples/translation). In our Fourierformer models, we choose $\phi(x) = x^2$ and the initialization $R_{\text{init}} = 1.0$.

### B Background

#### B.1 Kernel Density Estimation

Kernel density estimation (KDE) is the application of kernel smoothing for probability density estimation, i.e., a non-parametric method to estimate the probability density function of a random variable based on kernels as weights. Let $(x_1, x_2, \ldots, x_n)$ be i.i.d. samples drawn from some univariate distribution with an unknown density $f$ at any given point $x$. We are interested in estimating the shape of this function $f$. Its kernel density estimator is

$$
\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right),
$$

where $K$ is the kernel and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript $h$ is called the scaled kernel and defined as $K_h(x) = 1/h K(x/h)$.

#### B.2 Nonparametric Kernel Regression

Kernel regression is a nonparametric technique to estimate the conditional expectation of a random variable. The objective is to find a non-linear relation between a pair of random variables $X$ and $Y$. In any nonparametric regression, the conditional expectation of a variable $Y$ relative to a variable $X$ may be written:

$$
E(Y|X) = m(X),
$$

where $m$ is an unknown function.

**Nadaraya–Watson kernel regression** Nadaraya–Watson kernel regression estimates $m$ as a locally weighted average, using a kernel as a weighting function. The Nadaraya–Watson estimator is given by

$$
\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)},
$$

where $K_h$ is a scaled kernel with a bandwidth $h$.  

---

3 Implementation available at [link](https://github.com/thuml/Flowformer).
B.3 Fourier Integral Theorem

The Fourier integral theorem [92, 7] has been used in nonparametric mode clustering, deconvolution problem, and generative modeling [33]. It is a combination of Fourier transform and Fourier inverse transform. In particular, for any function $p \in \mathbb{L}_1(\mathbb{R}^D)$, the Fourier integral theorem is given by

$$ p(\mathbf{x}) = \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} \cos(s^T(x - y))p(y)dyds $$

$$ = \frac{1}{(2\pi)^D} \lim_{R \to \infty} \int_{\mathbb{R}^D} \int_{[-R,R]^D} \cos(s^T(x - y))p(y)dyds $$

$$ = \frac{1}{\pi^D} \lim_{R \to \infty} \int_{\mathbb{R}^D} \prod_{j=1}^D \sin(R(x_j - y_j)) \left(\frac{x_j - y_j}{D}\right) p(y)dy, \quad (20) $$

where $\mathbf{x} = (x_1, \ldots, x_D)$, $\mathbf{y} = (y_1, \ldots, y_D)$, $\mathbf{s} = (s_1, \ldots, s_D)$, and $R$ is the radius. Here, the first equality in equation (20) is due to

$$ \lim_{R \to \infty} \int_{[-R,R]^D} \cos(s^T(x - y))ds = \int_{\mathbb{R}^D} \cos(s^T(x - y))ds $$

and the final equality in equation (20) is due to

$$ \int_{[-R,R]^D} \cos(s^T(x - y))ds = \prod_{j=1}^D \sin(R(x_j - y_j)) \left(\frac{x_j - y_j}{D}\right) $$

for all $\mathbf{y} \in \mathbb{R}^D$. Equation (20) suggests that $p_R(\mathbf{x}) := \frac{1}{\pi^D} \int_{\mathbb{R}^D} \prod_{j=1}^D \sin(R(y_j - x_j)) \left(\frac{y_j - x_j}{D}\right) p(y)dy$ can be used as an estimator of the function $p$.

C Additional Theoretical Results

In this section, we provide additional theoretical results for generalized Fourier density estimator in Appendix C.1 and for generalized Fourier nonparametric regression estimator in Appendix C.2.

C.1 Generalized Fourier density estimator

We now establish the MISE rate of $p_{N,R}^\phi$ in equation (12) when $\phi(z) = z^l$ and $l \in \{1, 2\}$. We consider the following tail bounds on the Fourier transform of the true density function $p$ as follows.

**Definition 3** (1) We say that $p$ is supersmooth of order $\alpha$ if we have universal constants $C_1$ and $C_2$ such that the following inequalities hold for almost surely $\mathbf{x} \in \mathbb{R}^D$:

$$ |\widehat{p}(\mathbf{x})| \leq C_1 \exp\left(-C_2 \left(\sum_{j=1}^D |x_j|^\alpha\right)\right). $$

Here, $\widehat{p}$ denotes the Fourier transform of the function $p$.

(2) The function $p$ is ordinary smooth of order $\beta$ if there exists universal constant $c$ such that the following inequality holds for almost surely $\mathbf{x} \in \mathbb{R}^D$:

$$ |\widehat{p}(\mathbf{x})| \leq c \cdot \prod_{j=1}^D \frac{1}{(1 + |x_j|^\beta)}. $$

The notions of supersmoothness and ordinary smoothness had been used widely in deconvolution problems [28] and density estimation problems [20, 82, 33]. The supersmooth condition is satisfied when the function $p$ is Gaussian distribution or Cauchy distribution while the ordinary smooth condition is satisfied when the function $p$ is Laplace distribution and Beta distribution.

Based on the smoothness conditions in Definition 3, we have the following result regarding the mean-square integrated error (MISE) of the function generalized Fourier density estimator (12) (see equation (13) for a definition of MISE) when $\phi(z) = z^l$ and $l \in \{1, 2\}$.
Theorem 3 (a) When \( \phi(z) = z \), the following holds:

- (Supersmooth setting) If the true density function \( p \) is supersmooth function of order \( \alpha \) for some \( \alpha > 0 \), then there exists universal constants \( C_1, C_2, \) and \( C_3 \) such that as long as \( R \geq C_1 \) we have
  \[
  \text{MISE}(\hat{p}_{N,R}^\phi) \leq \hat{C}_2 \left( R^{\max\{1-\alpha,0\}} \exp(-\hat{C}_3 R^\alpha) + \frac{R^D}{N} \right).
  \]

- (Ordinary smooth setting) If the true density function \( p \) is ordinary smooth function of order \( \beta \) for some \( \beta > 1 \), then there exists universal constants \( \hat{c} \) such that
  \[
  \text{MISE}(\hat{p}_{N,R}^\phi) \leq \hat{c} \left( R^{-\beta+1} + \frac{R^D}{N} \right).
  \]

(b) When \( \phi(z) = z^2 \), the following holds

- (Supersmooth setting) If the true density function \( p \) is supersmooth function of order \( \alpha \) for some \( \alpha > 0 \), then there exists universal constants \( C_1', C_2' \) such that as long as \( R \geq C_1' \) we have
  \[
  \text{MISE}(\hat{p}_{N,R}^\phi) \leq \hat{C}_2' \left( \frac{1}{R^2} + \frac{R^D}{N} \right).
  \]

- (Ordinary smooth setting) If the true density function \( p \) is ordinary smooth function of order \( \beta \) for some \( \beta > 3 \), then there exists universal constants \( c' \) such that
  \[
  \text{MISE}(\hat{p}_{N,R}^\phi) \leq c' \left( \frac{1}{R^2} + \frac{R^D}{N} \right).
  \]

Proof of Theorem 3 is in Appendix D.2. A few comments with the results of Theorem 3 are in order.

When \( \phi(z) = z \): As part (a) of Theorem 3 indicates, when the function \( p \) is supersmooth, by choosing the radius \( R \) to balance the bias and variance, we have the optimal \( R \) as \( R = \left( \frac{\log(N)}{C_3} \right)^{1/\alpha} \) and the MISE rate of the generalized Fourier density estimator \( \hat{p}_{N,R}^\phi \) becomes \( O\left( \frac{\log(N)^{D/\alpha}}{N} \right) \). It indicates that, the MISE rate of \( \hat{p}_{N,R}^\phi \) is parametric when the function \( p \) is supersmooth. On the other hand, when the function \( p \) is ordinary smooth, the optimal \( R \) becomes \( R = O\left( N^{\frac{\beta-1}{\beta+1}} \right) \) and the MISE rate becomes \( O\left( N^{-\frac{\beta}{\beta+1}} \right) \). It is slower than the MISE rate when the function \( p \) is supersmooth.

When \( \phi(z) = z^2 \): The results of part (b) of Theorem 3 demonstrate that the upper bounds for the MISE rate of the generalized Fourier density estimator \( \hat{p}_{N,R}^\phi \) is similar for both the supersmooth and ordinary smooth settings. The optimal radius \( R = O\left( N^{\frac{1}{D+2}} \right) \) and the MISE rate of the estimator is \( O\left( N^{-\frac{2}{D+2}} \right) \).

C.2 Generalized Fourier nonparametric regression estimator

In this appendix, we provide additional result for the mean square error (MSE) rate of the generalized Fourier nonparametric regression estimator \( \hat{f}_{N,R} \) in equation (14) when \( \phi(z) = z \), namely, the setting of the Fourier integral theorem. The results when \( \phi(z) = z^l \) for \( l \in \{2, 3, 4, 5\} \) are left for the future work.

When \( \phi(z) = z \), the MSE rate of \( \hat{f}_{N,R} \) had been established in Theorem 9 of Ho et al. [33] when the function \( p \) is supersmooth function. Here, we restate that result for the completeness.

Theorem 4 Assume that the function \( p \) is supersmooth function of order \( \alpha \) for some \( \alpha > 0 \) and \( \sup_{k \in \mathbb{R}^d} |p(k)| < \infty \). Furthermore, we assume that the function \( f \) in the nonparametric regression model (3) is such that \( \sup_{k \in \mathbb{R}^d} |f^2(k)p(k)| < \infty \) and

\[
|\hat{f}\hat{p}(t)| \leq C_1 Q(|t_1|, |t_2|, \ldots, |t_D|) \exp \left( -C_2 \left( \sum_{j=1}^D |t_j|^\alpha \right) \right),
\]
where \( \hat{f}_p(t) \) is the Fourier transform of the function \( f, p \), \( C_1 \) and \( C_2 \) are some universal constants, and \( Q(|t_1|, |t_2|, \ldots, |t_D|) \) is some polynomial function of \( |t_1|, \ldots, |t_D| \) with non-negative coefficients. Then, we can find universal constants \( C_3, C_4, C_5 \) such that as long as \( R \geq C_3 \) we have

\[
\mathbb{E} \left[ (f_{N,R}(k) - f(k))^2 \right] \leq C_4 \frac{R^{\max\{2\deg(Q) + 2 - 2\alpha, 0\}}}{p^2(k) J(R)} \exp \left( -2C_2 R^\alpha \right) + \frac{\|f(k) + C_5 R^\alpha\|_N}{p^2(k) J(R)},
\]

where \( \deg(Q) \) denotes the degree of the polynomial function \( Q \), \( J(R) = 1 - \frac{R^{\max\{2 - 2\alpha, 0\}} \exp(-2C_2 R^\alpha)}{p^2(k) J(R)} \).

Proof of Theorem 4 is similar to the proof of Theorem 9 of Ho et al. [33]; therefore, it is omitted. The result of Theorem 4 indicates that the optimal radius \( R = \left( \frac{\log(N)}{2C_2} \right)^{1/\alpha} \) and the MSE rate of the generalized Fourier nonparametric regression estimator \( f_{N,R} \) is \( \mathcal{O}(\frac{\log(N)^{D/\alpha}}{N}) \).

D Proofs

In this Appendix, we provide proofs for key results in the paper and in Appendix C.

D.1 Proof of Theorem 1

Recall that, \( k_1, k_2, \ldots, k_N \in \mathbb{R}^D \) are i.i.d. samples from the density function \( p \). In equation (12), the generalized Fourier density estimator of \( p_0 \) is given by:

\[
p_{N,R}^\phi(k) = \frac{R^D}{N A^D} \sum_{i=1}^{N} \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right),
\]

where \( A = \int_R \phi \left( \frac{\sin(z)}{z} \right) dz \), \( k_i = (k_{i1}, \ldots, k_{iD}) \), and \( k = (k_1, \ldots, k_D) \). Direct calculation demonstrates that

\[
\mathbb{E}[p_{N,R}^\phi(k)] = \frac{R^D}{A^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - y_j))}{R(k_j - y_j)} \right) p(y) dy
\]

\[
= \frac{1}{A^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) p \left( k - \frac{y}{R} \right) dy.
\]

An application of Taylor expansion up to the \( m \)-th order indicates that

\[
p \left( k - \frac{y}{R} \right) = \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{m+1} \alpha!} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{m+1} p}{\partial k^\alpha}(k) + \tilde{R}(k, y),
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_d), |\alpha| = \sum_{j=1}^{D} \alpha_j \), and \( \tilde{R}(k, y) \) is Taylor remainder admitting the following form:

\[
\tilde{R}(k, y) = \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1} \beta!} \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_{0}^{1} (1 - t)^{m} \frac{\partial^{m+1} p}{\partial k^\beta} \left( k - \frac{ty}{R} \right) dt.
\]

Plugging equations (22) and (23) into equation (21), we find that

\[
\mathbb{E}[p_{N,R}^\phi(k)]
\]

\[
= p(k) + \frac{1}{A^D} \sum_{1 \leq |\alpha| \leq m} \frac{1}{R^{m+1} \alpha!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{m+1} p}{\partial k^\alpha}(k) dy
\]

\[
+ \frac{1}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1} \beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_{0}^{1} (1 - t)^{m} \frac{\partial^{m+1} p_0}{\partial k^\beta} \left( k - \frac{ty}{R} \right) dy dt.
\]
According to the hypothesis that \( \int_R \phi \left( \frac{\sin(z)}{z} \right) z^j dz = 0 \) for all \( 1 \leq j \leq m \), we obtain that
\[
\int_R \prod_{j=1}^D \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^D (-y_j)^{\alpha_j} \frac{\partial^{m+1} p}{\partial k^m} (k) dy = 0
\]
for any \( \alpha = (\alpha_1, \ldots, \alpha_d) \) such that \( 1 \leq |\alpha| \leq m \). Collecting the above results, we arrive at
\[
|\mathbb{E}[p_{N,R}^\phi(k)] - p(k)| = \left| \frac{1}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1} |\beta|!} \int_R \prod_{j=1}^D \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^D |y_j|^{\beta_j} dy \int_0^1 (1-t)^m \frac{\partial^{m+1} p}{\partial k^m} (k - ty) dy dt \right|
\]
\[
\leq \frac{1}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1} |\beta|!} \int_R \prod_{j=1}^D \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^D |y_j|^{\beta_j} \int_0^1 (1-t)^m \left| \frac{\partial^{m+1} p}{\partial k^m} (k - ty) \right| dy dt.
\]
Since the function \( p \in C^{m+1}(\mathbb{R}^D) \), we can find positive constant \( M \) such that \( \| \frac{\partial^{m+1} p}{\partial k^m} (k) \| \leq M \) for all \( \beta = (\beta_1, \ldots, \beta_D) \) such that \( |\beta| = m + 1 \). Therefore, we find that
\[
|\mathbb{E}[p_{N,R}^\phi(k)] - p(k)| \leq \frac{M}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1} |\beta|!} \int_R \prod_{j=1}^D \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^D |y_j|^{\beta_j} dy \int_0^1 (1-t)^m dt
\]
\[
= \frac{M}{A^D} \sum_{|\beta|=m+1} \frac{1}{R^{m+1} |\beta|!} \int_R \prod_{j=1}^D \phi \left( \frac{\sin(y_j)}{y_j} \right) \prod_{j=1}^D |y_j|^{\beta_j} dy.
\]
For any \( \beta = (\beta_1, \ldots, \beta_D) \) such that \( |\beta| = m + 1 \), an application of the AM-GM inequality indicates that \( \prod_{j=1}^D |y_j|^{\beta_j} \leq m! \sum_{j=1}^D |y_j|^{m+1} \). Hence, putting these results together leads to
\[
|\mathbb{E}[p_{N,R}^\phi(k)] - p(k)| \leq \frac{M m}{A^D R^{m+1}} \sum_{|\beta|=m+1} \frac{1}{|\beta|!} \int_R \prod_{j=1}^D \phi \left( \frac{\sin(y_j)}{y_j} \right) \left( \sum_{j=1}^D |y_j|^{m+1} \right) dy.
\]
From the hypothesis, we have \( \int_R \left| \phi \left( \frac{\sin(z)}{z} \right) \right| |z|^{m+1} dz < \infty \). As a consequence, we can find a universal constant \( C \) depending on \( A \) and \( d \) such that
\[
|\mathbb{E}[p_{N,R}^\phi(k)] - p(k)| \leq \frac{C}{R^{m+1}}
\]
for all \( k \in \mathbb{R}^D \).

**Bounding the Variance:** We now move to bound the variance of \( p_{N,R}^\phi(k) \). Indeed, direct computation indicates that
\[
\text{Var}[p_{N,R}^\phi(k)] = \frac{R^{2D}}{n A^2 D^2} \text{Var} \left[ \prod_{j=1}^D \phi \left( \frac{\sin(R(k_j - K_j))}{R(x_j - K_j)} \right) \right]
\]
\[
\leq \frac{R^{2D}}{n A^2 D^2} \mathbb{E} \left[ \prod_{j=1}^D \phi^2 \left( \frac{\sin(R(k_j - K_j))}{R(k_j - K_j)} \right) \right]
\]
\[
= \frac{R^D}{n A^2 D^2} \int_R \prod_{j=1}^D \phi^2 \left( \frac{\sin(y_j)}{y_j} \right) p(k - y) R dy \leq \frac{R^D}{N A^2 D^2} \int_R \prod_{j=1}^D \phi^2 \left( \frac{\sin(y_j)}{y_j} \right) dy
\]
where the expectation and the expectation are taken with respect to \( K = (K_1, \ldots, K_d) \sim p \). As \( \int_R \phi^2 \left( \frac{\sin(z)}{z} \right) dz < \infty \), there exists a universal constant \( C' \) depending on \( A \) and \( D \) such that
\[
\text{Var}[p_{N,R}^\phi(k)] \leq \frac{C'R^D}{N}.
\]
As a consequence, we obtain the conclusion of the theorem.
We first consider the setting when $b$.

Under this setting, direct computation indicates that

\[ \int_{\mathbb{R}^D} \left[ (\hat{p}_{N,R}^\phi(k) - p(k))^2 \right] dk = \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \left[ \hat{p}_{N,R}^\phi(t) - \hat{p}(t) \right]^2 dt, \quad (24) \]

where $\hat{p}_{N,R}^\phi$ and $\hat{p}$ are respectively the Fourier transforms of $p_{N,R}$ and $p$. From the definition of generalized Fourier density estimator $p_{N,R}^\phi$ in equation (12), it is clear that

\[ \hat{p}_{N,R}^\phi(t) = \frac{1}{N} \sum_{i=1}^N \exp(it^\top k_i) \prod_{j=1}^D K_R(t_j), \]

for any $t = (t_1, \ldots, t_D) \in \mathbb{R}^D$ where we define $K_R(y) := \frac{1}{\pi} \int_{\mathbb{R}} R\phi \left( \frac{\sin(yx)}{y} \right) \exp(iyx) dy$ for any $y \in \mathbb{R}$. To ease the presentation, we denote $\hat{K}_R(t) := \prod_{j=1}^D K_R(t_j)$ and $\varphi_N(t) = \frac{1}{N} \sum_{i=1}^N \exp(it^\top k_i)$ for any $t = (t_1, t_2, \ldots, t_D) \in \mathbb{R}^D$. Based on these notations, we can rewrite

\[ \hat{p}_{N,R}^\phi(t) = \varphi_N(t) \hat{K}_R(t) \]

Direct calculation shows that $E_k \varphi_N(t) = \hat{p}(t)$ for any $t \in \mathbb{R}^D$ where $k^n := (k_1, \ldots, k_n)$. Furthermore, we have

\[ E_k \varphi_N(t) = \varphi_N(-t) = \frac{1}{N} + \frac{(N-1)}{N} E_k \exp(it^\top k) \exp(-it^\top k) \]

Collecting the above results, we have the following equations:

\[ E_k \int_{\mathbb{R}^D} \left[ \hat{p}_{N,R}^\phi(t) - \hat{p}(t) \right]^2 dt = E_k \int_{\mathbb{R}^D} \left[ \varphi_N(t) \hat{K}_R(t) - \hat{p}(t) \right]^2 dt \]

\[ = E_k \int_{\mathbb{R}^D} \left[ (\varphi_N(t) - \hat{p}(t)) \hat{K}_R(t) - \hat{p}(t)(1 - \hat{K}_R(t)) \right]^2 dt \]

\[ = \int_{\mathbb{R}^D} E_k \left[ (\varphi_N(t) - \hat{p}(t))^2 \right] \hat{K}_R^2(t) + \hat{p}^2(t)(1 - \hat{K}_R(t))^2 dt \]

\[ = \int_{\mathbb{R}^D} \hat{p}^2(t)(1 - \hat{K}_R(t))^2 dt + \frac{1}{N} \int_{\mathbb{R}^D} (1 - |\hat{p}(t)|^2) \hat{K}_R^2(t) dt. \quad (25) \]

Combining the results from equations (24) and (25), we find that

\[ \text{MISE}(p_{N,R}^\phi) = E_k \int_{\mathbb{R}^D} \left[ (p_{N,R}^\phi(k) - p(k))^2 \right] dk \]

\[ = \frac{1}{(2\pi)^D} \left( \int_{\mathbb{R}^D} \hat{p}^2(t)(1 - \hat{K}_R(t))^2 dt + \frac{1}{N} \int_{\mathbb{R}^D} (1 - |\hat{p}(t)|^2) \hat{K}_R^2(t) dt \right). \quad (26) \]

**D.2.1 When $\phi(z) = z$**

We first consider the setting when $\phi(z) = z$, namely, the setting of the Fourier integral theorem. Under this setting, direct computation indicates that

\[ \hat{K}_R(t) = \prod_{i=1}^D 1_{\{|t_i| \leq R\}}. \]

Given the smoothness assumptions on the function $p$, we have two settings on that function.
Supersmooth setting of the function $p$: When the function $p$ is supersmooth density, we have

$$|\tilde{p}(t)| \leq C_1 \exp \left( -C_2 \left( \sum_{j=1}^{D} |t_j|^\alpha \right) \right),$$

where $C_1$ and $C_2$ are some universal constants. Therefore, we find that

$$\int_{\mathbb{R}^D} \tilde{p}^2(t)(1 - \tilde{K}_R(t))^2 dt = \int_{\mathbb{R}^D \setminus [-R,R]^D} \tilde{p}^2(t)dt \leq C_1 \int_{\mathbb{R}^D \setminus [-R,R]^D} \exp \left( -C_2 \left( \sum_{j=1}^{D} |t_j|^\alpha \right) \right) dt$$

$$\leq C_1 \sum_{i=1}^{D} \int_{B_i} \exp \left( -C_2 \left( \sum_{j=1}^{D} |t_j|^\alpha \right) \right) dt,$$

(27)

where $B_i := \{ t \in \mathbb{R}^D : |t_i| \geq R \}$. We now proceed to bound $\int_{B_i} \exp \left( -C_2 \left( \sum_{j=1}^{D} |t_j|^\alpha \right) \right) dt$ for all $i \in [D]$. Indeed, we have that

$$\int_{B_i} \exp \left( -C_2 \left( \sum_{j=1}^{D} |t_j|^\alpha \right) \right) dt = \left( \int_{R} \exp(-C_2|x|^\alpha)dx \right)^{D-1} \cdot \int_{|x| \geq R} \exp(-C_2|x|^\alpha)dx$$

$$= \frac{C_2 \alpha^{D-1}}{2C_2 \Gamma(1/\alpha)} \cdot \int_{|x| \geq R} \exp(-C_2|x|^\alpha)dx.$$

When $\alpha \geq 1$, we have that

$$\int_{R} \exp(-C_2x^\alpha)dx \leq \int_{R} x^{\alpha-1} \exp(-C_2x^\alpha)dx = \exp(-C_2R^\alpha)/(C_2\alpha).$$

When $\alpha \in (0, 1)$, then we find that

$$\int_{R} \exp(-C_2x^\alpha)dx = \int_{R} x^{1-\alpha} x^{\alpha-1} \exp(-C_2x^\alpha)dx$$

$$\leq \frac{R^{1-\alpha} \exp(-C_2R^\alpha)}{C_2\alpha} + \frac{1 - \alpha}{C_2\alpha R^\alpha} \int_{R} \exp(-C_2x^\alpha)dx,$$

When the $R$ is such that $R^\alpha \geq \frac{2(1-\alpha)}{C_2\alpha}$, the above inequality becomes

$$\int_{R} \exp(-C_2x^\alpha)dx \leq \frac{2R^{1-\alpha} \exp(-C_2R^\alpha)}{C_2\alpha}.$$

Collecting the above results, we arrive at

$$\int_{|x| \geq R} \exp(-C_2|x|^\alpha)dx \leq \frac{4R^{\max\{1-\alpha,0\}}}{C_2\alpha} \exp(-C_2R^\alpha).$$

(28)

Plugging the inequality (28) into the inequality (31), there exists universal constant $C_3$ depending on $\alpha$ and $D$ such that

$$\int_{\mathbb{R}^D} \tilde{p}^2(t)(1 - \tilde{K}_R(t))^2 dt \leq C_3 R^{\max\{1-\alpha,0\}} \exp(-C_1R^\alpha).$$

(29)

On the other hand, we also have

$$\frac{1}{N} \int_{\mathbb{R}^D} (1 - |\tilde{p}(t)|^2) \tilde{K}^2_R(t) dt \leq \frac{1}{N} \int_{\mathbb{R}^D} \tilde{K}^2_R(t) dt \leq \frac{R^D}{N}.$$

(30)

Combining the results from equations (29) and (30), we obtain that

$$\text{MISE}(P_{N,R}^\alpha) \leq C_4 \left( R^{\max\{1-\alpha,0\}} \exp(-C_1R^\alpha) + \frac{R^D}{N} \right).$$
As a consequence, we obtain the conclusion of Theorem 3 under the supersmooth setting of the function \( p \) and \( \phi(z) = z \).

**Ordinary smooth setting of the function** \( p \): The proof of Theorem 3 when the function \( p \) is ordinary smooth also proceeds in the similar fashion as that when \( p \) is supersmooth. In particular, we have

\[
\int_{\mathbb{R}^D} \tilde{p}^2(t)(1 - \bar{K}_R(t))^2 dt \leq c \sum_{i=1}^{D} \int_{B_i} \prod_{j=1}^{D} \frac{1}{(1 + |t_j|^\beta)} dt, \tag{31}
\]

where \( B_i := \{ t \in \mathbb{R}^D : |t_i| \geq R \} \). By simple algebra, we obtain that

\[
\int_{B_i} \prod_{j=1}^{D} \frac{1}{(1 + |t_j|^\beta)} dt = \left( \int_{\mathbb{R}} \frac{1}{1 + |x|^\beta} dx \right)^D \cdot \int_{|x| \geq R} \frac{1}{1 + |x|^\beta} dx \leq \left( \int_{\mathbb{R}} \frac{1}{1 + |x|^\beta} dx \right)^D \frac{2}{\beta - 1} R^{-\beta + 1}.
\]

Putting the above results together leads to

\[
\int_{\mathbb{R}^D} \tilde{p}^2(t)(1 - \bar{K}_R(t))^2 dt \leq c_1 R^{-\beta + 1}, \tag{32}
\]

where \( c_1 \) is some universal constant.

Similar to the supersmooth setting, we also can bound the variance \( \frac{1}{N} \int_{\mathbb{R}^D} (1 - |\hat{p}(t)|^2)K_R^2(t) dt \) under the ordinary smooth setting as follows:

\[
\frac{1}{N} \int_{\mathbb{R}^D} (1 - |\hat{p}(t)|^2)K_R^2(t) dt \leq \frac{R^D}{N}. \tag{33}
\]

Combining the results from equations (32) and (23), we obtain that

\[
\text{MISE}(p_{N,R}) \leq c_2 \left( R^{-\beta + 1} + \frac{R^D}{N} \right),
\]

where \( c_2 \) is a universal constant. As a consequence, we obtain the conclusion of Theorem 3 under the ordinary smooth setting of the function \( p \) and \( \phi(z) = z \).

**D.2.2 When** \( \phi(z) = z^2 \)

When \( \phi(z) = z^2 \), which corresponds to the Féjer integral setting, we find that

\[
\bar{K}_R(t) = \frac{1}{2D} \prod_{i=1}^{d} \left( 2 - \left| \frac{t_i}{R} \right| \right) \mathbf{1}_{\{|t_i| \leq 2R\}}.
\]

Given the formulation of the function \( \bar{K}_R \), we first bound \( \frac{1}{N} \int_{\mathbb{R}^D} (1 - |\hat{p}(t)|^2)\bar{K}_R^2(t) dt \). Indeed, direct calculation shows that

\[
\frac{1}{N} \int_{\mathbb{R}^D} (1 - |\hat{p}(t)|^2)\bar{K}_R^2(t) dt \leq \frac{1}{N} \int_{\mathbb{R}^D} \bar{K}_R^2(t) dt = \frac{1}{N^2} \left( \int_{|x| \leq 2R} \left( 2 - \frac{|x|}{R} \right) dx \right)^D = \frac{2^D R^D}{N}. \tag{34}
\]

Now, we proceed to upper bound \( \int_{\mathbb{R}^D} \tilde{p}^2(t)(1 - \bar{K}_R(t))^2 dt \). We have two settings of the function \( p \).

**Supersmooth setting of the function** \( p \): Given the above formulation of the function \( \bar{K}_R \), we have

\[
\int_{\mathbb{R}^D} \tilde{p}^2(t)(1 - \bar{K}_R(t))^2 dt = \int_{\mathbb{R}^D \setminus [-2R,2R]^D} \tilde{p}^2(t) dt
\]

\[
+ \int_{[-2R,2R]^D} \tilde{p}^2(t) \left( 1 - \prod_{i=1}^{D} \left( 1 - \frac{|t_i|}{2R} \right) \right)^2 dt. \tag{35}
\]
By using the similar argument as when \( \phi(x) = x \), when \( p \) is supersmooth function, we obtain that
\[
\int_{B^D \setminus [-2R,2R]^D} \tilde{p}^2(t) \, dt \leq C'_1 R^{\max\{1-\alpha,0\}} \exp(-C'_2 R^\alpha),
\] (36)
where \( C'_1 \) and \( C'_2 \) are universal constants. On the other hand, we have
\[
\int_{[-2R,2R]^D} \tilde{p}^2(t) \left( 1 - \prod_{i=1}^D \left( 1 - \frac{|t_i|}{2R} \right) \right)^2 \, dt
\leq C_1 \int_{[-2R,2R]^D} \exp \left( -C_2 \left( \sum_{j=1}^D |t_j|^{\alpha} \right) \right) \left( 1 - \prod_{i=1}^D \left( 1 - \frac{|t_i|}{2R} \right) \right)^2 \, dt
\leq C_1 \sum_{i_1, \ldots, i_m} \int_{[-2R,2R]^D} \exp \left( -C_2 \left( \sum_{j=1}^D |t_j|^{\alpha} \right) \right) \left( \prod_{i=1}^m \left( 1 - \frac{|t_{i_i}|}{2R} \right) \right)^2 \, dt,
\] (37)
where \( C_1 \) is some universal constant. Here, \( i_1, \ldots, i_m \) in the sum satisfy that they are pairwise different and \( 1 \leq i_1, \ldots, i_m \leq D \). Now, simple calculations indicate that
\[
\int_{[-2R,2R]^D} \exp \left( -C_2 \left( \sum_{j=1}^D |t_j|^{\alpha} \right) \right) \left( \prod_{i=1}^m \left( 1 - \frac{|t_{i_i}|}{2R} \right) \right)^2 \, dt \leq \frac{1}{R^{2m}} \int_{R^D} \exp \left( -C_2 \left( \sum_{j=1}^D |t_j|^{\alpha} \right) \right) \prod_{i=1}^m \left( 1 - \frac{|t_{i_i}|}{2R} \right)^2 \, dt \leq \frac{C_2}{R^{2m}},
\] (38)
where \( C_2 \) is some universal constant. Combining the results from equations (37) and (38), there exists universal constant \( C_3 \) depending on \( D \) such that
\[
\int_{[-2R,2R]^D} \tilde{p}^2(t) \left( 1 - \prod_{i=1}^D \left( 1 - \frac{|t_i|}{2R} \right) \right)^2 \, dt \leq \frac{C_3}{R^2}.
\] (39)
Plugging the inequalities (36) and (39) to equation (35) leads to the following bound
\[
\int_{R^D} \tilde{p}^2(t)(1 - \bar{K}_R(t))^2 \, dt \leq C'_1 R^{\max\{1-\alpha,0\}} \exp(-C'_2 R^\alpha) + \frac{C_3}{R^2} \leq \frac{C_4}{R^2}.
\] (40)
Combining the results from equations (34) and (40), we have
\[
\text{MISE}(\hat{p}^\phi_{N,R}) \leq \bar{C}_5 \left( \frac{1}{R^2} + \frac{R^D}{N} \right).
\]
As a consequence, we obtain the conclusion of Theorem 3 when \( \phi(z) = z^2 \) and the function \( p \) is supersmooth function.

**Ordinary smooth setting of the function \( p \):** Using similar proof argument as that of the supersmooth setting of the function \( p \), as \( \beta > 3 \), we find that
\[
\int_{R^D} \tilde{p}^2(t)(1 - \bar{K}_R(t))^2 \, dt \leq \frac{c}{R^{3-1}} + \int_{[-2R,2R]^D} \tilde{p}^2(t) \left( 1 - \prod_{i=1}^D \left( 1 - \frac{|t_i|}{2R} \right) \right)^2 \, dt
\leq \frac{c}{R^{3-1}} + \frac{c_1}{R^2} \leq \frac{c_2}{R^2},
\] (41)
where \( c, c_1, c_2 \) are universal constants. Combining the inequalities (34) and (41), we obtain the conclusion of Theorem 3 under the ordinary smooth setting of the function \( p \) and \( \phi(z) = z^2 \).

**D.3 Proof of Theorem 2**

Our proof strategy is to first bound the bias of \( f_{N,R}(k) \) and then establish an upper bound for the variance of \( f_{N,R}(k) \) for each \( k \in \mathbb{R}^D \).
D.3.1 Upper bound on the bias

Recall that in equation (14), we define $f_{N,R}(k)$ as follows:

$$f_{N,R}(k) := \frac{\sum_{i=1}^{N} \nu_i \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right)}{\sum_{i=1}^{N} \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right)} = \frac{a_{N,R}(k)}{p^\phi_{N,R}(k)},$$

where $p^\phi_{N,R}(k)$ is generalized Fourier density estimator in equation (12) while $a_{N,R}(k)$ is defined as follows:

$$a_{N,R}(k) := \frac{R^D}{nA^D} \sum_{i=1}^{N} \nu_i \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right).$$

Simple algebra leads to

$$f_{N,R}(k) - f(k) = \frac{a_{N,R}(k) - f(k)p^\phi_{N,R}(k)}{p(k)} + \frac{(f_{N,R}(k) - f(k))(p(k) - p^\phi_{N,R}(k))}{p(k)}. \quad (42)$$

Therefore, via an application of Cauchy-Schwarz inequality we obtain that

$$(\mathbb{E}[f_{N,R}(k)] - f(k))^2 \leq 2 \left( \frac{\mathbb{E}[a_{N,R}(k) - f(k)p^\phi_{N,R}(k)]}{p^2(k)} \right)^2 + 2 \frac{\mathbb{E}[(f_{N,R}(k) - f(k))(p(k) - p^\phi_{N,R}(k))]^2}{p^2(k)} \leq 2 \frac{\mathbb{E}[(f_{N,R}(k) - f(k))^2]}{p^2(k)} + 2 \frac{\mathbb{E}[(p(k) - p^\phi_{N,R}(k))^2]}{p^2(k)},$$

where the second inequality is due to the standard inequality $\mathbb{E}^2(XY) \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$ for all the random variables $X, Y$.

According to the assumptions of Theorem 2 and the result of Theorem 1, we have

$$\mathbb{E}[(p(k) - p^\phi_{N,R}(k))^2] \leq \frac{C_1}{R^{2(m+1)}} + \frac{C_2 R^D}{N^2}, \quad (44)$$

where $C_1$ and $C_2$ are some universal constants in Theorem 1.

Now, we proceed to bound $|\mathbb{E}[a_{N,R}(k) - f(k)p^\phi_{N,R}(k)]|$. Direct calculation demonstrates that

$$\mathbb{E}[a_{N,R}(k)] = \frac{R^D}{A^D} \int_{R^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \frac{p(y) f(y) dy}{y_j}$$

$$= \frac{1}{A^D} \int_{R^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) p \left( \frac{k - y}{R} \right) f \left( \frac{k - y}{R} \right) dy. \quad (45)$$

An application of Taylor expansion up to the $m$-th order indicates that

$$p \left( \frac{k - y}{R} \right) = \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{\alpha_1 + \alpha_2 + \ldots + \alpha_D}} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{\alpha_1 + \alpha_2 + \ldots + \alpha_D}}{\partial k^{\alpha_1 + \alpha_2 + \ldots + \alpha_D}} (k) + \hat{R}_1(k, y),$$

$$f \left( \frac{k - y}{R} \right) = \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{\alpha_1 + \alpha_2 + \ldots + \alpha_D}} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{\alpha_1 + \alpha_2 + \ldots + \alpha_D}}{\partial k^{\alpha_1 + \alpha_2 + \ldots + \alpha_D}} (k) + \hat{R}_2(k, y), \quad (46)$$

where $\alpha = (\alpha_1, \ldots, \alpha_D), |\alpha| = \sum_{j=1}^{d} \alpha_j$, and $\hat{R}_1(k, y), \hat{R}_2(k, y)$ are Taylor remainders admitting the following forms:

$$\hat{R}_1(k, y) = \sum_{|\beta|=m+1}^{m+1} \frac{m+1}{R^{m+1+|\beta|}} \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_{0}^{1} (1-t)^m \frac{\partial^{m+1}}{\partial k^{\beta}} \left( k - t \frac{y}{R} \right) dt,$$

$$\hat{R}_2(k, y) = \sum_{|\beta|=m+1}^{m+1} \frac{m+1}{R^{m+1+|\beta|}} \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_{0}^{1} (1-t)^m \frac{\partial^{m+1} f}{\partial k^{\beta}} \left( x - t \frac{y}{R} \right) dt. \quad (47)$$
Combining equations (46) and (47), we obtain that

\[
p \left( k - \frac{y}{R} \right) f \left( k - \frac{y}{R} \right) = \sum_{0 \leq |\alpha|, |\beta| \leq m} \frac{1}{R^{n+|\beta|}} \prod_{j=1}^{D} (-y_j)^{\alpha_j+\beta_j} \frac{\partial^{n|\beta|}}{\partial k^{|\alpha|}} p(k) \frac{\partial^{|\beta|} f}{\partial \kappa^{|\beta|}} (k) \]

\[
+ \left( \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{n|\alpha|}} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{n|\alpha|}}{\partial k^{|\alpha|}} p(k) \right) \bar{R}_2(k,y) \]

\[
+ \left( \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{n|\alpha|}} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|} f}{\partial \kappa^{|\alpha|}} \right) \bar{R}_1(k,y) + \bar{R}_1(k,y) \bar{R}_2(k,y).
\]

As we have \( \int_{\mathbb{R}} \phi \left( \frac{\sin(z)}{z} \right) z^j dz = 0 \) for all \( 1 \leq j \leq m \), plugging the equation in the above display to equation (45) leads to

\[
E[a_{n,R}(k)] = f(k)E[p_{N,R}(k)] + B_1 + B_2 + B_3 + B_4,
\]

where \( B_1, B_2, B_3, B_4 \) are defined as follows:

\[
B_1 = \frac{1}{\mathcal{A}^D} \sum_{m+1 \leq |\alpha|, |\beta| \leq 2m} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \frac{1}{R^{n+|\beta|}} \prod_{j=1}^{D} (-y_j)^{\alpha_j+\beta_j} \frac{\partial^{n|\beta|}}{\partial k^{|\alpha|}} p(k) \frac{\partial^{|\beta|} f}{\partial \kappa^{|\beta|}} (k) dy,
\]

\[
B_2 = \frac{1}{\mathcal{A}^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \left( \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{n|\alpha|}} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{n|\alpha|}}{\partial k^{|\alpha|}} p(k) \right) \bar{R}_2(k,y) dy,
\]

\[
B_3 = \frac{1}{\mathcal{A}^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \left( \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{n|\alpha|}} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|}}{\partial k^{|\alpha|}} f(k) \right) \bar{R}_1(k,y) dy,
\]

\[
B_4 = \frac{1}{\mathcal{A}^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) \bar{R}_1(k,y) \bar{R}_2(k,y) dy.
\]

Since we have \( \int_{\mathbb{R}} \left| \phi \left( \frac{\sin(z)}{z} \right) \right| |z|^j dz < \infty \) for any \( m+1 \leq j \leq 2m+2 \) and \( p_0, f \in \mathcal{C}^{m+1}(\mathbb{R}^d) \), we find that as long as \( R \geq \bar{c} \) for some given constant \( \bar{c} \)

\[
|B_1| \leq \frac{1}{\mathcal{A}^D} \sum_{m+1 \leq |\alpha|, |\beta| \leq 2m} \frac{1}{R^{n+|\beta|}} \prod_{j=1}^{D} \left| \phi \left( \frac{\sin(y_j)}{y_j} \right) \right| \prod_{j=1}^{D} |y_j|^{\alpha_j+\beta_j} \frac{\partial^{n|\beta|}}{\partial k^{|\alpha|}} p(k) \frac{\partial^{|\beta|} f}{\partial \kappa^{|\beta|}} \lesssim \frac{c_1}{R^{m+1}},
\]

where \( c_1 \) is some universal constant depending on \( A, D, \) and \( \bar{c} \). Furthermore, we find that

\[
|B_2| \leq \frac{1}{\mathcal{A}^D} \sum_{0 \leq |\alpha| \leq m} \frac{m+1}{R^{n+|\beta|}} \prod_{j=1}^{D} \left| \phi \left( \frac{\sin(y_j)}{y_j} \right) \right| \prod_{j=1}^{D} |y_j|^{\alpha_j} \int_0^1 (1-t)^m \frac{\partial f}{\partial \kappa} dt \lesssim \frac{c_2}{R^{m+1}},
\]

where \( c_2 \) is some universal constant depending on \( A, d, \) and \( \bar{c} \). Similarly, we also can demonstrate that \( B_3 \leq c_3/R^{m+1} \) and \( B_4 \leq c_4/R^{(m+1)} ) \) for some universal constants \( c_3 \) and \( c_4 \). Putting the above results together, we arrive at the following bound:

\[
E \left[ a_{n,R}(k) - f(k) \right] p_{N,R}(k) \right] \leq \frac{c'}{R^{m+1}}.
\]

Plugging the results from equations (44) and (48) to equation (43), we obtain that

\[
(E[f_{N,R}(k)] - f(k))^2 \leq \frac{2\epsilon'^2}{p^2(k)R^{2(m+1)}} + \frac{2\mathbb{E}[(f_{N,R}(k) - f(k))^2]}{p^2(k)} \left( \frac{C_1}{R^{2(m+1)}} + \frac{C_2R^D}{N} \right).
\]
D.3.2 Upper bound on the variance

Now, we study the variance of \( f_{N,R}(k) \). By taking variance both sides of the equation (42), we obtain that

\[
\text{var}(f_{N,R}(k)) = \text{var}\left( \frac{a_{N,R}(k) - f(k)p^\phi_{N,R}(k)}{p(k)} + \frac{(f_{N,R}(k) - f(k))(p(k) - p^\phi_{N,R}(k))}{p(k)} \right)
\]

\[
\leq \frac{2}{p^2(k)} \left( \mathbb{E}\left[ \left( \frac{a_{N,R}(k) - f(k)p^\phi_{N,R}(k)}{p(k)} \right)^2 \right]_{T_1} + \mathbb{E}\left[ (f_{N,R}(k) - f(k))^2(p(k) - p^\phi_{N,R}(k))^2 \right]_{T_2} \right).
\]

**Upper bound of \( T_2 \):** To upper bound \( T_2 \), we utilize the following lemma.

**Lemma 1** Assume that the function \( \phi \) and \( \rho_0 \) satisfy the assumptions of Theorem 1. Furthermore, \( \phi(z) \leq C \) as long as \( |z| \leq 1 \) for some universal constant \( C \). Then, for almost all \( k \in \mathbb{R}^D \), there exist universal constants \( C' \) such that

\[
\mathbb{P}\left( \left| p^\phi_{N,R}(k) - p(k) \right| \geq C' \left( \frac{1}{R^{m+1}} + \sqrt{\frac{RD \log(2/\delta)}{N}} \right) \right) \leq \delta.
\]

Proof of Lemma 1 is given in Appendix D.4. Now given the result of Lemma 1, we denote \( B \) as the event such that

\[
\left| p^\phi_{N,R}(k) - p(k) \right| \leq C' \left( \frac{1}{R^{m+1}} + \sqrt{\frac{RD \log(2/\delta)}{N}} \right)
\]

where \( C' \) is a universal constant in Lemma 1. Then, we obtain \( \mathbb{P}(B) \geq 1 - \delta \). Hence, we have the following bound for \( T_2 \):

\[
T_2 = \mathbb{E}\left[ (f_{N,R}(k) - f(k))^2(p(k) - p^\phi_{N,R}(k))^2 \right]_{B} \mathbb{P}(B) + \mathbb{E}\left[ (f_{N,R}(k) - f(k))^2(p(k) - p^\phi_{N,R}(k))^2 \right]_{B^c} \mathbb{P}(B^c)
\]

\[
\leq 2c' \mathbb{E}\left[ (f_{N,R}(k) - f(k))^2 \right] \left( \frac{1}{R^{2(m+1)}} + \frac{RD \log(2/\delta)}{N} + \delta \left( p^2(k) + \frac{C^2 RD^2}{A^2} \right) \right),
\]

where \( c' \) is some universal constant and the final inequality is based on the inequalities: \( \mathbb{P}(B^c) \leq \delta \) and \( (p(k) - p^\phi_{N,R}(k))^2 \leq 2(p^2(k) + (p^\phi_{N,R}(k))^2) \leq 2 \left( p^2(k) + \frac{C^2 RD^2}{A^2} \right) \) where \( C \) is a universal constant such that \( \phi(z) \leq C \) when \( |z| \leq 1 \). By choosing \( \delta \) such that \( \delta = \frac{R^D}{N\log^2(NR) + C^2 RD^2} \), we obtain that

\[
T_2 \leq c'' \mathbb{E}\left[ (f_{N,R}(k) - f(k))^2 \right] \left( \frac{1}{R^{2(m+1)}} + \frac{RD \log(NR)}{N} \right),
\]

for some universal constant \( c'' \) when \( R \) is sufficiently large.

**Upper bound of \( T_1 \):** As \( v_i = f(k_i) + \epsilon_i \) for all \( i \in [N] \), direct calculation shows that

\[
T_1 = \mathbb{E}\left[ \left( \frac{R^D}{NA^2} \sum_{i=1}^{N} (f(k_i) - f(k)) \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right) \right) \right]_{T_1} + \mathbb{E}\left[ \left( \frac{R^D}{NA^2} \sum_{i=1}^{N} \epsilon_i \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right) \right)^2 \right]_{T_2}\]
An application of Cauchy-Schwarz inequality leads to

\[
T_1 \leq 2E \left[ \left( \frac{R^D}{NA^D} \sum_{i=1}^{N} (f(k_i) - f(k)) \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right) \right)^2 \right] + 2E \left[ \left( \frac{1}{N\pi^D} \sum_{i=1}^{N} \epsilon_i \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right) \right)^2 \right] = 2(S_1 + S_2).
\]

Since we have \( E \left[ \left( \frac{1}{N} \sum_{i=1}^{N} Z_i \right)^2 \right] \leq \frac{1}{N} E[Z_1^2] + E[Z_1] \) for any i.i.d. samples \( Z_1, \ldots, Z_N \), we obtain that

\[
S_1 \leq \frac{R^D}{NA^D} E \left[ (f(X) - f(k))^2 \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right] + \frac{R^D}{A^D} E^2 \left[ (f(X) - f(k))^2 \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right],
\]

where the outer expectation is taken with respect to \( X = (X_1, \ldots, X_d) \sim p \). From the result in equation (48), we have

\[
\frac{R^D}{A^D} E^2 \left[ (f(X) - f(k))^2 \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right] = E^2 \left[ a_{N,R}(k) - f(k)p_{N,R}(k) \right] \leq \frac{c'}{R^{2(m+1)}},
\]

where \( c' \) is some universal constant. In addition, an application of Cauchy-Schwarz inequality leads to

\[
\frac{R^D}{NA^D} E \left[ (f(X) - f(k))^2 \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right]
\leq \frac{2R^D}{NA^D} E \left[ (f^2(X) + f^2(k)) \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right]
\leq \frac{2R^D}{NA^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) f^2(k, y) \prod_{j=1}^{D} \phi \left( \frac{\sin(y_j)}{y_j} \right) dy.
\]

Since we have \( \int_{\mathbb{R}} \phi^2(\sin(z)/z)dz < \infty \), it indicates that we can find a universal constant \( c'' \) such that

\[
\frac{R^D}{NA^D} E \left[ (f(X) - f(k))^2 \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right] \leq c'' \frac{R^D}{NA^D} \left( \|f^2 \times p\|_\infty + f^2(k) \right).
\]

Putting the above results together, we obtain that

\[
S_1 \leq \frac{c'}{R^{2(m+1)}} + \frac{c'' R^D}{NA^D} \left( \|f^2 \times p\|_\infty + f^2(k) \right).
\]

Similarly, since \( E(\epsilon_i) = 0 \) and \( \text{var}(\epsilon_i) = \sigma^2 \) for all \( i \in [N] \), we have

\[
S_2 = \frac{\sigma^2 R^D}{NA^D} E \left[ \prod_{j=1}^{D} \phi \left( \frac{\sin(R(k_j - X_j))}{R(k_j - X_j)} \right) \right] \leq \frac{c'''' \sigma^2 R^D}{NA^D} \left( \|p\|_\infty R^D \right),
\]

(53)
where $c''$ is some universal constant. Combining the results from equation (52) and equation (53), we find that

$$T_1 \leq C \left( \frac{\|f^2 \times p\|_\infty + f^2(k) + \sigma^2 \|p\|_\infty R^D}{N} + \frac{1}{R^{2m+1}} \right),$$

where $C$ is some universal constant. Plugging the bounds of $T_1$ and $T_2$ from equations (51) and (54) into equation (50), when $R \geq C'$ where $C'$ is some universal constant, we have

$$\text{var}(f_{N,R}(k)) \leq \frac{C_1'}{p^2(k)} \mathbb{E} \left[ (f_{N,R}(k) - f(k))^2 \right] \left( \frac{1}{R^{2(m+1)}} + \frac{R^D \log(NR)}{N} \right) + \frac{C_2'}{p^2(k)} \left( \frac{f(k) + C_3' R^D}{N} + \frac{1}{R^{2(m+1)}} \right),$$

where $C_1', C_2', C_3'$ are some universal constants. Combining the results with bias and variance in equations (49) and (55), we obtain the following bound:

$$\mathbb{E} \left[ (f_{N,R}(k) - f(k))^2 \right] \leq \frac{2(c')^2}{p^2(k) R^{2(m+1)}} + \frac{2 \mathbb{E} \left[ (f_{N,R}(k) - f(k))^2 \right]}{p^2(k)} \left( \frac{C_1}{R^{2(m+1)}} + \frac{C_2 R^D}{N} \right) + \frac{C_1'}{p^2(k)} \mathbb{E} \left[ (f_{N,R}(k) - f(k))^2 \right] \left( \frac{1}{R^{2(m+1)}} + \frac{R^D \log(NR)}{N} \right) + \frac{C_2'}{p^2(k)} \left( \frac{f(k) + C_3' R^D}{N} + \frac{1}{R^{2(m+1)}} \right).$$

As a consequence, we obtain the conclusion of the theorem.

**D.4 Proof of Lemma 1**

Invoking triangle inequality, we obtain that

$$\left| p_{N,R}^\phi(k) - p(k) \right| \leq \left| p_{N,R}^\phi(k) - \mathbb{E} \left[ p_{N,R}^\phi(k) \right] \right| + \left| \mathbb{E} \left[ p_{N,R}^\phi(k) \right] - p(k) \right|. \quad (56)$$

If we denote $v_i = \frac{R^D}{A^D} \prod_{j=1}^D \phi \left( \frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})} \right)$ for all $i \in [N]$, then as $\sin(R(k_j - k_{ij}))/R(k_j - k_{ij}) \leq 1$ for all $j \in [D]$ we have $|v_i| \leq C^D R^D / A^D$ for all $i \in [N]$ where $C$ is the constant such that $\phi(z) \leq C$ when $|z| \leq 1$. Furthermore, from the proof of Theorem 1 we have $\text{var}(v_i) \leq C'R^D$ where $C' > 0$ is some universal constant. Given these bounds of $v_i$ and $\text{var}(v_i)$, for any $t \in (0, C'')$ Bernstein’s inequality shows that

$$\mathbb{P} \left( \frac{1}{N} \sum_{i=1}^N |v_i - \mathbb{E} |v_i| | \geq t \right) \leq 2 \exp \left( - \frac{Nt^2}{2C'R^D + 2C'R^D t/(3A^D)} \right).$$

By choosing $t = C \sqrt{R^D \log(2/\delta) / N}$, where $C$ is some universal constant, we find that

$$\mathbb{P} \left( \left| p_{N,R}^\phi(k) - \mathbb{E} \left[ p_{N,R}^\phi(k) \right] \right| \geq t \right) = \mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^N v_i - \mathbb{E} |v_i| \right| \geq t \right) \leq \delta. \quad (57)$$

From the result of Theorem 1, there exists universal constant $c$ such that

$$\mathbb{E} \left[ p_{N,R}^\phi(k) \right] - p(k) \leq c/R^{m+1}. \quad (58)$$

Plugging the bounds (57) and (58) into the triangle inequality (56), we obtain the conclusion of the lemma.

**E Additional Experimental Results**

**E.1 Effect of $\phi$**

Using the WikiText-103 language modeling as a case study, we analyze the effect of $\phi(x)$ on the performance of FourierFormer. In particular, we set $\phi(x) = x^k$ and compare the performance of FourierFormer for $k = 1, 2, 3, 4$ and 6. We keep other settings the same as in our experiments in
Table 7. Ablation study on how the choice of $\phi(x) = x^k$ influences the performance of FourierFormer. Odd values of $k$ cause training to diverge. For even values of $k$, greater $k$ yields better perplexity (PPL), but the improvement is small for $k > 4$. Other choices of $\phi$ such as $\phi(x) = |x|$, ReLU$(x)$, and sigmoid$(x)$ yield worse results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Valid PPL</th>
<th>Test PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline dot-product (small)</td>
<td>33.15</td>
<td>34.29</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = x^2$ (small)</td>
<td>32.09</td>
<td>33.10</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = x^4$ (small)</td>
<td>31.86</td>
<td>32.85</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = x^8$ (small)</td>
<td>31.84</td>
<td>32.81</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = x$ (small)</td>
<td>not converge</td>
<td>not converge</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = x^3$ (small)</td>
<td>not converge</td>
<td>not converge</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) =</td>
<td>x</td>
<td>$ (small)</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = \text{ReLU}(x)$ (small)</td>
<td>33.37</td>
<td>35.01</td>
</tr>
<tr>
<td>FourierFormer, $\phi(x) = \text{sigmoid}(x)$ (small)</td>
<td>not converge</td>
<td>not converge</td>
</tr>
</tbody>
</table>

Table 8. Ablation study on how the initialization of $R$ influences the performance of FourierFormer. When $R$ is initialized to a too small or too big value, the PPL of the trained FourierFormer is reduced. $R_{\text{init}} = 1, 2, 3$ yield the best results. FourierFormer with learnable vectors $R$ yields better results than Fourierformer of the same setting using learnable scalars $R$ with the cost of increasing the number of parameters in the model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Valid PPL</th>
<th>Test PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline dot-product (small)</td>
<td>33.15</td>
<td>34.29</td>
</tr>
<tr>
<td>FourierFormer, $R_{\text{init}} = 0.1$ (small)</td>
<td>32.04</td>
<td>33.01</td>
</tr>
<tr>
<td>FourierFormer, $R_{\text{init}} = 1.0$ (small)</td>
<td>31.89</td>
<td>32.87</td>
</tr>
<tr>
<td>FourierFormer, $R_{\text{init}} = 2.0$ (small)</td>
<td>31.86</td>
<td>32.85</td>
</tr>
<tr>
<td>FourierFormer, $R_{\text{init}} = 3.0$ (small)</td>
<td>31.90</td>
<td>32.88</td>
</tr>
<tr>
<td>FourierFormer, $R_{\text{init}} = 4.0$ (small)</td>
<td>32.58</td>
<td>33.65</td>
</tr>
<tr>
<td>FourierFormer, $R_{\text{init}} = 2.0$ (small, $R$ is a vector)</td>
<td>31.82</td>
<td>32.80</td>
</tr>
</tbody>
</table>

Section 4.1. We summarize our results in Table 7. We observe that for odd values of $k$ such as $k = 1, 3$, the training diverges, confirming that negative density estimator cause instability in training FourierFormer (see Remark 3.1). For even values of $k$ such as $k = 2, 4, 6$, we observe that the greater value of $k$ results in better valid and test PPL. However, the gap between $k = 4$ and $k = 6$ is smaller compared to the gap between $k = 2$ and $k = 4$, suggesting that using $k > 4$ does not add much advantage in terms of accuracy. We have also studied other choices of $\phi$ that are nonnegative functions such as $\phi(x) = |x|$, ReLU$(x)$, and sigmoid$(x)$. Those functions yield worse results than $\phi(x) = x^{2m}$. We summarize these results in Table 7.

E.2 Effect of the Initialization of $R$

In this section, we study the effect of the initialization value of $R$ on the performance of FourierFormer when trained for the WikiText-103 language modeling and summarize our results in Table 8. Here we choose $R$ to be learnable scalars as in experiments described in our main text. Other settings are also the same as in our experiments in Section 4.1. We observe that when $R$ is initialized too small (e.g. $R_{\text{init}} = 0.1$) or too big (e.g. $R_{\text{init}} = 4$), the PPL of the trained FourierFormer decreases. $R_{\text{init}} = 1, 2, 3$ yield best results. We also study the performance of the FourierFormer when $R$ is chosen to be a learnable vector, $R = [R_1, \ldots, R_D]^\top$. We report our result in the last row of Table 8. FourierFormer with $R$ be learnable vectors achieves better PPLs than FourierFormer with $R$ be learnable scalars of the same setting. As we mentioned in Section A, this advantage comes with an increase in the number of parameters in the model. Finally, from our experiments, we observe that making $R$ a learnable parameter yields better PPLs than making $R$ a constant and selecting its value via a careful search.

E.3 Efficiency Analysis

We have included quantitative results on the runtime and GPU memory usage of the FourierFormer versus the baseline softmax transformer in Table 9.

E.4 Synthetic Examples for Density Estimation and Nonparametric Regression via The Generalized Fourier Integral Theorem

We empirically confirm Theorem 1 for density estimation and Theorem 2 for nonparametric regression using the Generalized Fourier Integral Theorem in this section. In Figure 1, we show that the generalized Fourier density estimator can approximate (A) 1-D and (B) 2-D Gaussian distribution
Table 9. Runtime and GPU memory usage of the FourierFormer vs. the baseline softmax transformer. Both models are trained for the WikiText-103 language modeling task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Runtime (Train) (miliseconds/sample)</th>
<th>GPU Memory (Train) (GB)</th>
<th>Runtime (Test) (miliseconds/sample)</th>
<th>GPU Memory (Test) (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline softmax (small)</td>
<td>5.41</td>
<td>1.43</td>
<td>1.53</td>
<td>0.94</td>
</tr>
<tr>
<td>FourierFormer (small)</td>
<td>6.00</td>
<td>1.43</td>
<td>1.70</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Figure 1. (A) 1-D and (B) 2-D Gaussian distributions and their estimated densities via Fourier Integral theorem.

Figure 2: Non-parametric regression via the Fourier Integral theorem.

with a dense covariance matrix well, which further verify Theorem 1. In Figure 2, we show that the generalized Fourier nonparametric regression estimator can approximate the function that maps from a random variable to another random variable, which further verify Theorem 2.

In particular, for the density estimation experiments, we sample 100000 data points from the 1-D and 2-D Gaussian distribution and estimate the density for 1000 uniformly sampled test points. The mean square errors (MSE) are $1.29 \times 10^{-5}$ and $2.42 \times 10^{-5}$ for the 1-D and 2-D case, respectively. For the non-parametric regression task, we build a training dataset with 90000 correlated normally distributed samples and choose a 3-degree polynomial as the ground truth function. The MSE between ground truth labels and predictions is 0.06.