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].[1] 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

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

We provide an ablation study for the effect of R and ϕ 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 scalar initialized to 1 and choose $\phi(x) = x^4$. Different attention units have different R.

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.

[1] Implementation available at https://github.com/IDSIA/lmtool-fwp.
[2] Implementation available at https://github.com/facebookresearch/deit.

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] [3] (for the PEMS-SF, SelfRegulationSCP2, UWaveGestureLibrary datasets) and [95] [4] (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 https://github.com/pytorch/fairseq/tree/main/examples/translation. In our Fourierformer models, we choose $\phi(x) = x^2$ and the initialization $R_{init} = 1.0$.

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 univarite 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

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{17}$$

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 $K_h(x) = 1/hK(x/h)$.

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:

$$E(Y|X) = m(X), \tag{18}$$

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

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}, \tag{19}$$

where K_h is a scaled kernel with a bandwidth h.

[3] Implementation available at https://github.com/thuml/Flowformer.

[4] Implementation available at https://github.com/gzerveas/mvts_transformer.

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

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} \cos(\boldsymbol{s}^\top(\boldsymbol{x}-\boldsymbol{y}))p(\boldsymbol{y})d\boldsymbol{y}d\boldsymbol{s} \\ &= \frac{1}{(2\pi)^D} \lim_{R\to\infty} \int_{\mathbb{R}^D} \int_{[-R,R]^D} \cos(\boldsymbol{s}^\top(\boldsymbol{x}-\boldsymbol{y}))p(\boldsymbol{y})d\boldsymbol{y}d\boldsymbol{s} \\ &= \frac{1}{\pi^D} \lim_{R\to\infty} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \frac{\sin(R(x_j-y_j))}{(x_j-y_j)} p(\boldsymbol{y})d\boldsymbol{y}, \end{aligned} \tag{20}$$

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

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

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

$$\int_{[-R,R]^D} \cos(\boldsymbol{s}^\top(\boldsymbol{x}-\boldsymbol{y}))d\boldsymbol{s} = \prod_{j=1}^{D} \frac{\sin(R(x_j-y_j))}{(x_j-y_j)}$$

for all $\boldsymbol{y} \in \mathbb{R}^D$. Equation (20) suggests that $p_R(\boldsymbol{x}) := \frac{1}{\pi^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \frac{\sin(R(y_j-x_j))}{(y_j-x_j)} p(\boldsymbol{y})d\boldsymbol{y}$ 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 $p^{\phi}_{N,R}$ 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 α if we have universal constants C_1 and C_2 such that the following inequalities hold for almost surely $x \in \mathbb{R}^D$:*

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

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

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

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

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 α for some $\alpha > 0$, then there exists universal constants $\bar{C}_1, \bar{C}_2$, and $\bar{C}_3$ such that as long as $R \geq \bar{C}_1$ we have*

$$\textit{MISE}(p^{\phi}_{N,R}) \leq \bar{C}_2 \left(R^{\max\{1-\alpha,0\}} \exp(-\bar{C}_3 R^{\alpha}) + \frac{R^D}{N} \right).$$

- *(Ordinary smooth setting) If the true density function p is ordinary smooth function of order β for some $\beta > 1$, then there exists universal constants $\bar{c}$ such that*

$$\textit{MISE}(p^{\phi}_{N,R}) \leq \bar{c} \left(R^{-\beta+1} + \frac{R^D}{N} \right).$$

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

- *(Supersmooth setting) If the true density function p is supersmooth function of order α for some $\alpha > 0$, then there exists universal constants C_1' and C_2' such that as long as $R \geq C_1'$ we have*

$$\textit{MISE}(p^{\phi}_{N,R}) \leq C_2' \left(\frac{1}{R^2} + \frac{R^D}{N} \right).$$

- *(Ordinary smooth setting) If the true density function p is ordinary smooth function of order β for some $\beta > 3$, then there exists universal constants c' such that*

$$\textit{MISE}(p^{\phi}_{N,R}) \leq c' \left(\frac{1}{R^2} + \frac{R^D}{N} \right).$$

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 $R = \left(\frac{\log(N)}{\bar{C}_3}\right)^{1/\alpha}$ and the MISE rate of the generalized Fourier density estimator $p^{\phi}_{N,R}$ becomes $\mathcal{O}\left(\frac{\log(N)^{D/\alpha}}{N}\right)$. It indicates that, the MISE rate of $p^{\phi}_{N,R}$ is parametric when the function p is supersmooth. On the other hand, when the function p is ordinary smooth, the optimal R becomes $R = \mathcal{O}(N^{\frac{1}{D+\beta-1}})$ and the MISE rate becomes $\mathcal{O}\left(N^{-\frac{\beta-1}{D+\beta-1}}\right)$. 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 $p^{\phi}_{N,R}$ is similar for both the supersmooth and ordinary smooth settings. The optimal radius $R = \mathcal{O}\left(N^{\frac{1}{D+2}}\right)$ and the MISE rate of the estimator is $\mathcal{O}\left(N^{-\frac{2}{D+2}}\right)$.

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 $f_{N,R}$ 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 $f_{N,R}$ 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 α for some $\alpha > 0$ and $\sup_{\boldsymbol{k} \in \mathbb{R}^D} |p(\boldsymbol{k})| < \infty$. Furthermore, we assume that the function f in the nonparametric regression model (3) is such that $\sup_{\boldsymbol{k} \in \mathbb{R}^D} |f^2(\boldsymbol{k}) p(\boldsymbol{k})| < \infty$ and*

$$|\widehat{f.p}(\boldsymbol{t})| \leq C_1 Q(|t_1|, |t_2|, \ldots, |t_D|) \exp\left(-C_2 \left(\sum_{j=1}^{D} |t_j|^{\alpha} \right) \right),$$

where $\widehat{f.p}(\boldsymbol{t})$ 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

$$\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right] \leq C_4 \frac{R^{\max\{2deg(Q)+2-2\alpha,0\}} \exp\left(-2C_2R^\alpha\right) + \frac{(f(\boldsymbol{k})+C_5)R^D}{N}}{p^2(\boldsymbol{k})\bar{J}(R)},$$

where $deg(Q)$ denotes the degree of the polynomial function Q, $\bar{J}(R) = 1 - \frac{R^{\max\{2-2\alpha,0\}} \exp(-2C_2R^\alpha) + \frac{R^D \log(NR)}{N}}{p^2(\boldsymbol{k})}$.

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 $\mathcal{O}\left(\frac{\log(N)^{D/\alpha}}{N}\right)$.

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, $\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_N \in \mathbb{R}^D$ are i.i.d. samples from the density function p. In equation (12), the generalized Fourier density estimator of p_0 is given by:

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

where $A = \int_{\mathbb{R}} \phi\left(\frac{\sin(z)}{z}\right) dz$, $\boldsymbol{k}_i = (k_{i1}, \ldots, k_{iD})$, and $\boldsymbol{k} = (k_1, \ldots, k_D)$. Direct calculation demonstrates that

$$\begin{aligned}
\mathbb{E}[p^\phi_{N,R}(\boldsymbol{k})] &= \frac{R^D}{A^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(R(k_j - y_j))}{R(k_j - y_j)}\right) p(\boldsymbol{y}) d\boldsymbol{y} \\
&= \frac{1}{A^D} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(y_j)}{y_j}\right) p\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) d\boldsymbol{y}. \qquad (21)
\end{aligned}$$

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

$$p\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) = \sum_{0 \leq |\alpha| \leq m} \frac{1}{R^{|\alpha|}\alpha!} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|} p}{\partial \boldsymbol{k}^\alpha}(\boldsymbol{k}) + \bar{R}(\boldsymbol{k}, \boldsymbol{y}), \qquad (22)$$

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

$$\bar{R}(\boldsymbol{k}, \boldsymbol{y}) = \sum_{|\beta| = m+1} \frac{m+1}{R^{m+1}\beta!} \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_0^1 (1-t)^m \frac{\partial^{m+1} p}{\partial \boldsymbol{k}^\beta}\left(\boldsymbol{k} - \frac{t\boldsymbol{y}}{R}\right) dt. \qquad (23)$$

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

$$\begin{aligned}
&\mathbb{E}[p^\phi_{N,R}(\boldsymbol{k})] \\
&= p(\boldsymbol{k}) + \frac{1}{A^D} \sum_{1 \leq |\alpha| \leq m} \frac{1}{R^{|\alpha|}\alpha!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(y_j)}{y_j}\right) \prod_{j=1}^{d} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|} p}{\partial \boldsymbol{k}^\alpha}(\boldsymbol{k}) d\boldsymbol{y} \\
&+ \frac{1}{A^D} \sum_{|\beta| = m+1} \frac{m+1}{R^{m+1}\beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(y_j)}{y_j}\right) \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_0^1 (1-t)^m \frac{\partial^{m+1} p_0}{\partial \boldsymbol{k}^\beta}\left(\boldsymbol{k} - \frac{t\boldsymbol{y}}{R}\right) d\boldsymbol{y} dt.
\end{aligned}$$

According to the hypothesis that $\int_{\mathbb{R}} \phi\left(\frac{\sin(z)}{z}\right) z^j dz = 0$ for all $1 \leq j \leq m$, we obtain that

$$\int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(y_j)}{y_j}\right) \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|} p}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k}) d\boldsymbol{y} = 0$$

for any $\alpha = (\alpha_1, \ldots, \alpha_d)$ such that $1 \leq |\alpha| \leq m$. Collecting the above results, we arrive at

$$\begin{aligned}
&|\mathbb{E}[p^{\phi}_{N,R}(\boldsymbol{k})] - p(\boldsymbol{k})| \\
&= \left| \frac{1}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1}\beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(y_j)}{y_j}\right) \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_0^1 (1-t)^m \frac{\partial^{m+1} p}{\partial \boldsymbol{k}^{\beta}}\left(\boldsymbol{k} - \frac{t\boldsymbol{y}}{R}\right) d\boldsymbol{y} dt \right| \\
&\leq \frac{1}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1}\beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \left|\phi\left(\frac{\sin(y_j)}{y_j}\right)\right| \prod_{j=1}^{D} |y_j|^{\beta_j} \int_0^1 (1-t)^m \left|\frac{\partial^{m+1} p}{\partial \boldsymbol{k}^{\beta}}\left(\boldsymbol{k} - \frac{t\boldsymbol{y}}{R}\right)\right| d\boldsymbol{y} dt.
\end{aligned}$$

Since the function $p \in \mathcal{C}^{m+1}(\mathbb{R}^D)$, we can find positive constant M such that $\|\frac{\partial^{m+1} p}{\partial \boldsymbol{k}^{\beta}}(\boldsymbol{k})\|_{\infty} \leq M$ for all $\beta = (\beta_1, \ldots, \beta_d)$ such that $|\beta| = m+1$. Therefore, we find that

$$\begin{aligned}
|\mathbb{E}[p^{\phi}_{N,R}(\boldsymbol{k})] - p(\boldsymbol{k})| &\leq \frac{M}{A^D} \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1}\beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \left|\phi\left(\frac{\sin(y_j)}{y_j}\right)\right| \prod_{j=1}^{D} |y_j|^{\beta_j} d\boldsymbol{y} \int_0^1 (1-t)^m dt \\
&= \frac{M}{A^D} \sum_{|\beta|=m+1} \frac{1}{R^{m+1}\beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \left|\phi\left(\frac{\sin(y_j)}{y_j}\right)\right| \prod_{j=1}^{D} |y_j|^{\beta_j} d\boldsymbol{y}.
\end{aligned}$$

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

$$|\mathbb{E}[p^{\phi}_{N,R}(\boldsymbol{k})] - p(\boldsymbol{k})| \leq \frac{Mm}{A^D R^{m+1}} \sum_{|\beta|=m+1} \frac{1}{\beta!} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \left|\phi\left(\frac{\sin(y_j)}{y_j}\right)\right| \left(\sum_{j=1}^{D} |y_j|^{m+1}\right) d\boldsymbol{y}.$$

From the hypothesis, we have $\int_{\mathbb{R}} \left|\phi\left(\frac{\sin(z)}{z}\right)\right| |z|^{m+1} dz < \infty$. As a consequence, we can find a universal constant C depending on A and d such that

$$|\mathbb{E}[p^{\phi}_{n,R}(\boldsymbol{k})] - p(\boldsymbol{k})| \leq \frac{C}{R^{m+1}}$$

for all $\boldsymbol{k} \in \mathbb{R}^D$.

Bounding the variance: We now move to bound the variance of $p^{\phi}_{N,R}(\boldsymbol{k})$. Indeed, direct computation indicates that

$$\begin{aligned}
\text{Var}[p^{\phi}_{N,R}(\boldsymbol{k})] &= \frac{R^{2D}}{nA^{2D}} \text{Var}\left[\prod_{j=1}^{D} \phi\left(\frac{\sin(R(k_j - K_{\cdot j}))}{R(x_j - K_{\cdot j})}\right)\right] \\
&\leq \frac{R^{2D}}{nA^{2D}} \mathbb{E}\left[\prod_{j=1}^{D} \phi^2\left(\frac{\sin(R(k_j - K_{\cdot j}))}{R(k_j - K_{\cdot j})}\right)\right] \\
&= \frac{R^D}{nA^{2D}} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi^2\left(\frac{\sin(y_j))}{y_j}\right) p\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) d\boldsymbol{y} \leq \frac{R^D \|p\|_{\infty}}{NA^{2D}} \int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi^2\left(\frac{\sin(y_j))}{y_j}\right) d\boldsymbol{y}
\end{aligned}$$

where the variance and the expectation are taken with respect to $K = (K_{\cdot 1}, \ldots, K_{\cdot d}) \sim p$. As $\int_{\mathbb{R}} \phi^2\left(\frac{\sin(z))}{z}\right) dz < \infty$, there exists a universal constant C' depending on A and D such that

$$\text{Var}[p^{\phi}_{N,R}(\boldsymbol{k})] \leq \frac{C' R^D}{N}.$$

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

D.2 Proof of Theorem 3

From the Plancherel theorem, we obtain that

$$\int_{\mathbb{R}^D} \left[(p^{\phi}_{N,R}(\boldsymbol{k}) - p(\boldsymbol{k})\right]^2 d\boldsymbol{k} = \frac{1}{(2\pi)^D} \int_{\mathbb{R}^D} \left[\widehat{p}^{\phi}_{N,R}(\boldsymbol{t}) - \widehat{p}(\boldsymbol{t})\right]^2 d\boldsymbol{t}, \tag{24}$$

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

$$\widehat{p}^{\phi}_{N,R}(t) = \frac{1}{N}\sum_{i=1}^{N} \exp(i\boldsymbol{t}^\top \boldsymbol{k}_i) \prod_{j=1}^{D} K_R(t_j),$$

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

$$\widehat{p}^{\phi}_{N,R}(\boldsymbol{t}) = \varphi_N(\boldsymbol{t})\bar{K}_R(\boldsymbol{t})$$

Direct calculation shows that $\mathbb{E}_{\boldsymbol{k}_1^N}[\varphi_N(\boldsymbol{t})] = \widehat{p}(\boldsymbol{t})$ for any $\boldsymbol{t} \in \mathbb{R}^D$ where $\boldsymbol{k}_1^N := (\boldsymbol{k}_1, \ldots, \boldsymbol{k}_n)$. Furthermore, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{k}_1^N}[|\varphi_N(\boldsymbol{t})|^2] = \mathbb{E}[\varphi_N(\boldsymbol{t})\varphi_N(-\boldsymbol{t})] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\exp(i\boldsymbol{t}^\top\boldsymbol{k}_i)\right)\left(\frac{1}{N}\sum_{i=1}^{N}\exp(-i\boldsymbol{t}^\top\boldsymbol{k}_i)\right)\right] \\
&= \frac{1}{N} + \frac{(N-1)}{N}\mathbb{E}\left[\exp(i\boldsymbol{t}^\top\boldsymbol{k})\exp(-i\boldsymbol{t}^\top\boldsymbol{k})\right] \\
&= \frac{1}{N} + \frac{(N-1)}{N}|\widehat{p}(\boldsymbol{t})|^2.
\end{aligned}$$

Collecting the above results, we have the following equations:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{k}_1^n}\left[\int_{\mathbb{R}^D}\left[\widehat{p}^{\phi}_{N,R}(\boldsymbol{t}) - \widehat{p}(\boldsymbol{t})\right]^2 d\boldsymbol{t}\right] &= \mathbb{E}_{\boldsymbol{k}_1^n}\left[\int_{\mathbb{R}^D}\left[\varphi_N(\boldsymbol{t})\bar{K}_R(\boldsymbol{t}) - \widehat{p}(\boldsymbol{t})\right]^2 d\boldsymbol{t}\right] \\
&= \mathbb{E}_{\boldsymbol{k}_1^n}\left[\int_{\mathbb{R}^D}\left[(\varphi_n(\boldsymbol{t}) - \widehat{p}(\boldsymbol{t}))\bar{K}_R(\boldsymbol{t}) - \widehat{p}(\boldsymbol{t})(1-\bar{K}_R(\boldsymbol{t}))\right]^2 d\boldsymbol{t}\right] \\
&= \int_{\mathbb{R}^D}\mathbb{E}_{\boldsymbol{k}_1^N}\left[(\varphi_N(\boldsymbol{t}) - \widehat{p}(\boldsymbol{t}))^2\right]\bar{K}_R^2(\boldsymbol{t}) + \widehat{p}^2(\boldsymbol{t})(1-\bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} \\
&= \int_{\mathbb{R}^D}\widehat{p}^2(\boldsymbol{t})(1-\bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} + \frac{1}{N}\int_{\mathbb{R}^D}(1-|\widehat{p}(\boldsymbol{t})|^2)\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t}.
\end{aligned} \tag{25}$$

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

$$\begin{aligned}
\text{MISE}(p^{\phi}_{N,R}) &= \mathbb{E}_{\boldsymbol{k}_1^N}\left[\int_{\mathbb{R}^D}\left[(p^{\phi}_{N,R}(\boldsymbol{k}) - p(\boldsymbol{k})\right]^2 d\boldsymbol{k}\right] \\
&= \frac{1}{(2\pi)^D}\left(\int_{\mathbb{R}^D}\widehat{p}^2(\boldsymbol{t})(1-\bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} + \frac{1}{N}\int_{\mathbb{R}^D}(1-|\widehat{p}(\boldsymbol{t})|^2)\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t}\right).
\end{aligned} \tag{26}$$

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

$$\bar{K}_R(\boldsymbol{t}) = \prod_{i=1}^{d}\mathbf{1}_{\{|t_i|\le R\}}.$$

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

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

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

$$\begin{aligned}\int_{\mathbb{R}^D}\widehat{p}^2(\boldsymbol{t})(1-\bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} = \int_{\mathbb{R}^D\setminus[-R,R]^D}\widehat{p}^2(\boldsymbol{t})d\boldsymbol{t} &\leq C_1\int_{\mathbb{R}^D\setminus[-R,R]^D}\exp\left(-C_2\left(\sum_{j=1}^{D}|t_j|^{\alpha}\right)\right)d\boldsymbol{t}\\ &\leq C_1\sum_{i=1}^{D}\int_{B_i}\exp\left(-C_2\left(\sum_{j=1}^{D}|t_j|^{\alpha}\right)\right)d\boldsymbol{t},\end{aligned} \tag{27}$$

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

$$\begin{aligned}\int_{B_i}\exp\left(-C_2\left(\sum_{j=1}^{D}|t_j|^{\alpha}\right)\right)d\boldsymbol{t} &= \left(\int_{\mathbb{R}}\exp(-C_2|x|^{\alpha})dx\right)^{D-1}\cdot\int_{|x|\geq R}\exp(-C_2|x|^{\alpha})dx\\ &= \frac{C_2\alpha^{D-1}}{(2C_2\Gamma(1/\alpha))^{D-1}}\cdot\int_{|x|\geq R}\exp(-C_2|x|^{\alpha})dx.\end{aligned}$$

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

$$\int_R^{\infty}\exp\left(-C_2x^{\alpha}\right)dx \leq \int_R^{\infty}x^{\alpha-1}\exp\left(-C_2x^{\alpha}\right)dx = \exp(-C_2R^{\alpha})/(C_2\alpha).$$

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

$$\begin{aligned}\int_R^{\infty}\exp(-C_2x^{\alpha})dx &= \int_R^{\infty}x^{1-\alpha}x^{\alpha-1}\exp(-C_2x^{\alpha})dx\\ &\leq \frac{R^{1-\alpha}\exp\left(-C_2R^{\alpha}\right)}{C_2\alpha}+\frac{1-\alpha}{C_2\alpha R^{\alpha}}\int_R^{\infty}\exp(-C_2x^{\alpha})dx,\end{aligned}$$

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

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

Collecting the above results, we arrive at

$$\int_{|x|\geq R}\exp(-C_2|x|^{\alpha})dx \leq \frac{4R^{\max\{1-\alpha,0\}}}{C_2\alpha}\exp(-C_2R^{\alpha}). \tag{28}$$

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

$$\int_{\mathbb{R}^D}\widehat{p}^2(\boldsymbol{t})(1-\bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} \leq C_3R^{\max\{1-\alpha,0\}}\exp(-C_1R^{\alpha}). \tag{29}$$

On the other hand, we also have

$$\frac{1}{N}\int_{\mathbb{R}^D}(1-|\widehat{p}(\boldsymbol{t})|^2)\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t} \leq \frac{1}{N}\int_{\mathbb{R}^D}\bar{K}_R^2(\boldsymbol{t}) \leq \frac{R^D}{N}. \tag{30}$$

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

$$\text{MISE}(p_{N,R}^{\phi}) \leq C_4\left(R^{\max\{1-\alpha,0\}}\exp(-C_1R^{\alpha})+\frac{R^D}{N}\right).$$

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

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

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

$$\begin{aligned}\int_{B_i} \prod_{j=1}^{D} \frac{1}{(1 + |t_j|^\beta)} d\boldsymbol{t} &= \left(\int_{\mathbb{R}} \frac{1}{1 + |x|^\beta} dx\right)^{D-1} \cdot \int_{|x| \geq R} \frac{1}{1 + |x|^\beta} dx \\ &\leq \left(\int_{\mathbb{R}} \frac{1}{1 + |x|^\beta} dx\right)^{D-1} \frac{2}{\beta - 1} R^{-\beta+1}.\end{aligned}$$

Putting the above results together leads to

$$\int_{\mathbb{R}^D} \widehat{p}^2(\boldsymbol{t})(1 - \bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} \leq c_1 R^{-\beta+1}, \tag{32}$$

where c_1 is some universal constant.

Similar to the supersmooth setting, we also can bound the variance $\frac{1}{N}\int_{\mathbb{R}^D}(1 - |\widehat{p}(\boldsymbol{t})|^2)\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t}$ under the ordinary smooth setting as follows:

$$\frac{1}{N}\int_{\mathbb{R}^D}(1 - |\widehat{p}(\boldsymbol{t})|^2)\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t} \leq \frac{R^D}{N}. \tag{33}$$

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

$$\text{MISE}(p^{\phi}_{N,R}) \leq c_2 \left(R^{-\beta+1} + \frac{R^D}{N}\right),$$

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

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

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

$$\begin{aligned}\frac{1}{N}\int_{\mathbb{R}^D}(1 - |\widehat{p}(\boldsymbol{t})|^2)\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t} \leq \frac{1}{N}\int_{\mathbb{R}^D}\bar{K}_R^2(\boldsymbol{t})d\boldsymbol{t} &= \frac{1}{N2^D}\left(\int_{|x| \leq 2R}\left(2 - \frac{|x|}{R}\right)dx\right)^D \\ &= \frac{2^D R^D}{N}.\end{aligned} \tag{34}$$

Now, we proceed to upper bound $\int_{\mathbb{R}^D} \widehat{p}^2(\boldsymbol{t})(1 - \bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t}$. 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

$$\begin{aligned}\int_{\mathbb{R}^D} \widehat{p}^2(\boldsymbol{t})(1 - \bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} &= \int_{\mathbb{R}^D \setminus [-2R, 2R]^D} \widehat{p}^2(\boldsymbol{t}) d\boldsymbol{t} \\ &\quad + \int_{[-2R, 2R]^D} \widehat{p}^2(\boldsymbol{t}) \left(1 - \prod_{i=1}^{D}\left(1 - \frac{|t_i|}{2R}\right)\right)^2 d\boldsymbol{t}.\end{aligned} \tag{35}$$

By using the similar argument as when $\phi(x) = x$, when p is supersmooth function, we obtain that

$$\int_{\mathbb{R}^D \setminus [-2R,2R]^D} \widehat{p}^2(\boldsymbol{t}) d\boldsymbol{t} \leq C_1' R^{\max\{1-\alpha,0\}} \exp(-C_2' R^\alpha), \tag{36}$$

where C_1' and C_2' are universal constants. On the other hand, we have

$$\begin{aligned}
&\int_{[-2R,2R]^D} \widehat{p}^2(\boldsymbol{t}) \left(1 - \prod_{i=1}^D \left(1 - \frac{|t_i|}{2R}\right)\right)^2 d\boldsymbol{t} \\
&\qquad \leq C_1 \int_{[-2R,2R]^D} \exp\left(-C_2 \left(\sum_{j=1}^D |t_j|^\alpha\right)\right) \left(1 - \prod_{i=1}^D \left(1 - \frac{|t_i|}{2R}\right)\right)^2 d\boldsymbol{t} \\
&\qquad \leq \bar{C}_1 \sum_{m=1}^D \sum_{i_1,\ldots,i_m} \int_{[-2R,2R]^D} \exp\left(-C_2 \left(\sum_{j=1}^D |t_j|^\alpha\right)\right) \frac{\prod_{l=1}^m t_{i_l}^2}{R^{2m}} d\boldsymbol{t},
\end{aligned} \tag{37}$$

where $\bar{C}_1$ 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

$$\begin{aligned}
&\int_{[-2R,2R]^D} \exp\left(-C_2 \left(\sum_{j=1}^D |t_j|^\alpha\right)\right) \frac{\prod_{l=1}^m t_{i_l}^2}{R^{2m}} d\boldsymbol{t} \leq \\
&\qquad \frac{1}{R^{2m}} \int_{\mathbb{R}^D} \exp\left(-C_2 \left(\sum_{j=1}^D |t_j|^\alpha\right)\right) \prod_{l=1}^m t_{i_l}^2 d\boldsymbol{t} \leq \frac{\bar{C}_2}{R^{2m}},
\end{aligned} \tag{38}$$

where $\bar{C}_2$ is some universal constant. Combining the results from equations (37) and (38), there exists universal constant $\bar{C}_3$ depending on D such that

$$\int_{[-2R,2R]^D} \widehat{p}^2(\boldsymbol{t}) \left(1 - \prod_{i=1}^D \left(1 - \frac{|t_i|}{2R}\right)\right)^2 d\boldsymbol{t} \leq \frac{\bar{C}_3}{R^2}. \tag{39}$$

Plugging the inequalities (36) and (39) to equation (35) leads to the following bound

$$\int_{\mathbb{R}^D} \widehat{p}^2(\boldsymbol{t})(1 - \bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} \leq C_1' R^{\max\{1-\alpha,0\}} \exp(-C_2' R^\alpha) + \frac{\bar{C}_3}{R^2} \leq \frac{\bar{C}_4}{R^2}. \tag{40}$$

Combining the results from equations (34) and (40), we have

$$\text{MISE}(p_{N,R}^\phi) \leq \bar{C}_5 \left(\frac{1}{R^2} + \frac{R^D}{N}\right).$$

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

$$\begin{aligned}
\int_{\mathbb{R}^D} \widehat{p}^2(\boldsymbol{t})(1 - \bar{K}_R(\boldsymbol{t}))^2 d\boldsymbol{t} &\leq \frac{c}{R^{\beta-1}} + \int_{[-2R,2R]^D} \widehat{p}^2(\boldsymbol{t}) \left(1 - \prod_{i=1}^D \left(1 - \frac{|t_i|}{2R}\right)\right)^2 d\boldsymbol{t} \\
&\leq \frac{c}{R^{\beta-1}} + \frac{c_1}{R^2} \leq \frac{c_2}{R^2},
\end{aligned} \tag{41}$$

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 $f_{N,R}(\boldsymbol{k})$ and then establish an upper bound for the variance of $f_{N,R}(\boldsymbol{k})$ for each $\boldsymbol{k} \in \mathbb{R}^D$.

D.3.1 Upper bound on the bias

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

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

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

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

Simple algebra leads to

$$f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}) = \frac{a_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k})p^{\phi}_{N,R}(\boldsymbol{k})}{p(\boldsymbol{k})} + \frac{(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))(p(\boldsymbol{k}) - p^{\phi}_{n,R}(\boldsymbol{k}))}{p(\boldsymbol{k})}. \tag{42}$$

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

$$\begin{aligned}
&\left(\mathbb{E}\left[f_{N,R}(\boldsymbol{k})\right] - f(\boldsymbol{k})\right)^2 \\
&\leq 2\frac{\left(\mathbb{E}\left[a_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k})p^{\phi}_{N,R}(\boldsymbol{k})\right]\right)^2}{p^2(\boldsymbol{k})} + 2\frac{\left(\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))\right]\right)^2}{p^2(\boldsymbol{k})} \\
&\leq 2\frac{\left(\mathbb{E}\left[a_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k})p^{\phi}_{N,R}(\boldsymbol{k})\right]\right)^2}{p^2(\boldsymbol{k})} + 2\frac{\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]\mathbb{E}\left[(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))^2\right]}{p^2(\boldsymbol{k})},
\end{aligned} \tag{43}$$

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

$$\mathbb{E}\left[(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))^2\right] \leq \frac{C_1}{R^{2(m+1)}} + \frac{C_2 R^D}{N}, \tag{44}$$

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

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

$$\begin{aligned}
\mathbb{E}\left[a_{N,R}(\boldsymbol{k})\right] &= \frac{R^D}{A^D}\int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(R(k_j-y_j))}{R(k_j-y_j)}\right) p(\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y} \\
&= \frac{1}{A^D}\int_{\mathbb{R}^D} \prod_{j=1}^{D} \phi\left(\frac{\sin(y_j)}{y_j}\right) p\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) f\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) d\boldsymbol{y}.
\end{aligned} \tag{45}$$

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

$$\begin{aligned}
p\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) &= \sum_{0\leq|\alpha|\leq m} \frac{1}{R^{|\alpha|}\alpha!} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|} p}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k}) + \bar{R}_1(\boldsymbol{k}, \boldsymbol{y}), \\
f\left(\boldsymbol{k} - \frac{\boldsymbol{y}}{R}\right) &= \sum_{0\leq|\alpha|\leq m} \frac{1}{R^{|\alpha|}\alpha!} \prod_{j=1}^{D} (-y_j)^{\alpha_j} \frac{\partial^{|\alpha|} f}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k}) + \bar{R}_2(\boldsymbol{k}, \boldsymbol{y}),
\end{aligned} \tag{46}$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha| = \sum_{j=1}^{d} \alpha_j$, and $\bar{R}_1(\boldsymbol{k}, \boldsymbol{y})$, $R_2(\boldsymbol{k}, \boldsymbol{y})$ are Taylor remainders admitting the following forms:

$$\begin{aligned}
\bar{R}_1(\boldsymbol{k}, \boldsymbol{y}) &= \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1}\beta!} \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_0^1 (1-t)^m \frac{\partial^{m+1} p}{\partial \boldsymbol{k}^{\beta}}\left(\boldsymbol{k} - \frac{t\boldsymbol{y}}{R}\right) dt, \\
\bar{R}_2(\boldsymbol{k}, \boldsymbol{y}) &= \sum_{|\beta|=m+1} \frac{m+1}{R^{m+1}\beta!} \prod_{j=1}^{D} (-y_j)^{\beta_j} \int_0^1 (1-t)^m \frac{\partial^{m+1} f}{\partial \boldsymbol{k}^{\beta}}\left(x - \frac{t\boldsymbol{y}}{R}\right) dt.
\end{aligned} \tag{47}$$

Combining equations (46) and (47), we obtain that

$$p\left(\boldsymbol{k}-\frac{\boldsymbol{y}}{R}\right) f\left(\boldsymbol{k}-\frac{\boldsymbol{y}}{R}\right) = \sum_{0\leq|\alpha|,|\beta|\leq m} \frac{1}{R^{|\alpha|+|\beta|}\alpha!\beta!}\prod_{j=1}^{D}(-y_j)^{\alpha_j+\beta_j}\frac{\partial^{|\alpha|}p}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k})\frac{\partial^{|\beta|}f}{\partial \boldsymbol{k}^{\beta}}(\boldsymbol{k}) + \left(\sum_{0\leq|\alpha|\leq m}\frac{1}{R^{|\alpha|}\alpha!}\prod_{j=1}^{D}(-y_j)^{\alpha_j}\frac{\partial^{|\alpha|}p}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k})\right)\bar{R}_2(\boldsymbol{k},\boldsymbol{y}) + \left(\sum_{0\leq|\alpha|\leq m}\frac{1}{R^{|\alpha|}\alpha!}\prod_{j=1}^{D}(-y_j)^{\alpha_j}\frac{\partial^{|\alpha|}f}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k})\right)\bar{R}_1(\boldsymbol{k},\boldsymbol{y}) + \bar{R}_1(\boldsymbol{k},\boldsymbol{y})\bar{R}_2(\boldsymbol{k},\boldsymbol{y}).$$

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

$$\mathbb{E}\left[a_{n,R}(\boldsymbol{k})\right] = f(\boldsymbol{k})\mathbb{E}\left[p^{\phi}_{N,R}(\boldsymbol{k})\right] + B_1 + B_2 + B_3 + B_4,$$

where B_1, B_2, B_3, B_4 are defined as follows:

$$B_1 = \frac{1}{A^D}\sum_{m+1\leq|\alpha|+|\beta|\leq 2m}\int_{\mathbb{R}^D}\prod_{j=1}^{D}\phi\left(\frac{\sin(y_j)}{y_j}\right)\frac{1}{R^{|\alpha|+|\beta|}\alpha!\beta!}\prod_{j=1}^{D}(-y_j)^{\alpha_j+\beta_j}\frac{\partial^{|\alpha|}p}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k})\frac{\partial^{|\beta|}f}{\partial \boldsymbol{k}^{\beta}}(\boldsymbol{k})d\boldsymbol{y},$$

$$B_2 = \frac{1}{A^D}\int_{\mathbb{R}^D}\prod_{j=1}^{D}\phi\left(\frac{\sin(y_j)}{y_j}\right)\left(\sum_{0\leq|\alpha|\leq m}\frac{1}{R^{|\alpha|}\alpha!}\prod_{j=1}^{D}(-y_j)^{\alpha_j}\frac{\partial^{|\alpha|}p_0}{\partial \boldsymbol{k}^{\alpha}}(\boldsymbol{k})\right)\bar{R}_2(\boldsymbol{k},\boldsymbol{y})d\boldsymbol{y},$$

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

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

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 $\bar{c}$

$$|B_1| \leq \frac{1}{A^D}\sum_{m+1\leq|\alpha|+|\beta|\leq 2m}\frac{1}{R^{|\alpha|+|\beta|}\alpha!\beta!}\int_{\mathbb{R}^D}\prod_{j=1}^{D}\left|\phi\left(\frac{\sin(y_j)}{y_j}\right)\right|\prod_{j=1}^{D}|y_j|^{\alpha_j+\beta_j}\|\frac{\partial^{|\alpha|}p}{\partial \boldsymbol{k}^{\alpha}}\|_{\infty}\|\frac{\partial^{|\beta|}f}{\partial \boldsymbol{k}^{\beta}}\|_{\infty} \leq \frac{c_1}{R^{m+1}},$$

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

$$|B_2| \leq \frac{1}{A^D}\sum_{0\leq|\alpha|\leq m,|\beta|=m+1}\frac{m+1}{R^{|\alpha|+m+1}\alpha!\beta!}\int_{\mathbb{R}^D}\prod_{j=1}^{D}\left|\phi\left(\frac{\sin(y_j)}{y_j}\right)\right|\prod_{j=1}^{D}|y_j|^{\alpha_j+\beta_j} \times\int_0^1(1-t)^m\|\frac{\partial^{m+1}f}{\partial \boldsymbol{k}^{\beta}}\|_{\infty}d\boldsymbol{y}dt \leq \frac{c_2}{R^{m+1}},$$

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 $B_4\leq c_4/R^{2(m+1)}$ for some universal constants c_3 and c_4. Putting the above results together, we arrive at the following bound:

$$\left|\mathbb{E}\left[a_{n,R}(\boldsymbol{k}) - f(\boldsymbol{k})p^{\phi}_{N,R}(\boldsymbol{k})\right]\right| \leq \frac{c'}{R^{m+1}}. \tag{48}$$

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

$$\left(\mathbb{E}\left[f_{N,R}(\boldsymbol{k})\right] - f(\boldsymbol{k})\right)^2 \leq \frac{2(c')^2}{p^2(\boldsymbol{k})R^{2(m+1)}} + \frac{2\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]}{p^2(\boldsymbol{k})}\left(\frac{C_1}{R^{2(m+1)}} + \frac{C_2R^D}{N}\right). \tag{49}$$

D.3.2 Upper bound on the variance

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

$$\begin{aligned}\mathrm{var}(f_{N,R}(\boldsymbol{k})) &= \mathrm{var}\left(\frac{a_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k})p^{\phi}_{N,R}(\boldsymbol{k})}{p(\boldsymbol{k})} + \frac{(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))}{p(\boldsymbol{k})}\right) \\ &\leq \frac{2}{p^2(\boldsymbol{k})}\left(\underbrace{\mathbb{E}\left[\left(a_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k})p^{\phi}_{N,R}(\boldsymbol{k})\right)^2\right]}_{T_1} + \underbrace{\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))^2\right]}_{T_2}\right).\end{aligned} \tag{50}$$

Upper bound of T_2: To upper bound T_2, we utilize the following lemma.

Lemma 1 *Assume that the function ϕ and p_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 $\boldsymbol{k} \in \mathbb{R}^D$, there exist universal constants C' such that*

$$\mathbb{P}\left(\left|p^{\phi}_{N,R}(\boldsymbol{k}) - p(\boldsymbol{k})\right| \geq C'\left(\frac{1}{R^{m+1}} + \sqrt{\frac{R^D\log(2/\delta)}{N}}\right)\right) \leq \delta.$$

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

$$\left|p^{\phi}_{N,R}(\boldsymbol{k}) - p(\boldsymbol{k})\right| \leq C'\left(\frac{1}{R^{m+1}} + \sqrt{\frac{R^D\log(2/\delta)}{N}}\right)$$

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

$$\begin{aligned}T_2 &= \mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))^2|B\right]\mathbb{P}(B) \\ &\qquad + \mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2(p(\boldsymbol{k}) - p^{\phi}_{N,R}(\boldsymbol{k}))^2|B^c\right]\mathbb{P}(B^c) \\ &\leq 2c'\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]\left(\frac{1}{R^{2(m+1)}} + \frac{R^D\log(2/\delta)}{N} + \delta\left(p^2(\boldsymbol{k}) + \frac{C^DR^{2D}}{A^D}\right)\right),\end{aligned}$$

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

$$T_2 \leq c''\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]\left(\frac{1}{R^{2(m+1)}} + \frac{R^D\log(NR)}{N}\right), \tag{51}$$

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

Upper bound of T_1: As $\mathbf{v}_i = f(\boldsymbol{k}_i) + \epsilon_i$ for all $i \in [N]$, direct calculation shows that

$$\begin{aligned}T_1 = \mathbb{E}\Bigg[\Bigg(&\frac{R^D}{NA^D}\sum_{i=1}^{N}(f(\boldsymbol{k}_i) - f(\boldsymbol{k}))\prod_{j=1}^{D}\phi\left(\frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})}\right) \\ &+ \frac{R^D}{NA^D}\sum_{i=1}^{N}\epsilon_i\prod_{j=1}^{D}\phi\left(\frac{\sin(R(k_j - k_{ij}))}{R(k_j - k_{ij})}\right)\Bigg)^2\Bigg].\end{aligned}$$

An application of Cauchy-Schwarz inequality leads to

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

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

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

where the outer expectation is taken with respect to $X=(X_{\cdot 1},\ldots,X_{\cdot d})\sim p$. From the result in equation (48), we have

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

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

$$\begin{aligned}
&\frac{R^{2D}}{NA^{2D}}\mathbb{E}\left[(f(X)-f(\boldsymbol{k}))^2\prod_{j=1}^{D}\phi^2\left(\frac{\sin(R(k_j-X_{\cdot j}))}{R(k_j-X_{\cdot j})}\right)\right]\\
&\quad\leq \frac{2R^{2D}}{NA^{2D}}\mathbb{E}\left[(f^2(X)+f^2(\boldsymbol{k}))\prod_{j=1}^{D}\phi^2\left(\frac{\sin(R(k_j-X_{\cdot j}))}{R(k_j-X_{\cdot j})}\right)\right]\\
&\quad= \frac{2R^{D}}{NA^{2D}}\int_{\mathbb{R}^D}\prod_{j=1}^{D}\phi^2\left(\frac{\sin(y_j))}{y_j}\right)\left(f^2\left(\boldsymbol{k}-\frac{\boldsymbol{y}}{R}\right)p\left(\boldsymbol{k}-\frac{\boldsymbol{y}}{R}\right)+f^2(\boldsymbol{k})\right)d\boldsymbol{y}\\
&\quad\leq \frac{2R^{D}(\|f^2\times p\|_\infty+f^2(\boldsymbol{k}))}{NA^{2D}}\int_{\mathbb{R}^D}\prod_{j=1}^{D}\phi^2\left(\frac{\sin(y_j))}{y_j}\right)dy.
\end{aligned}$$

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

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

Putting the above results together, we obtain that

$$S_1 \leq \frac{c'}{R^{2(m+1)}} + \frac{c''R^D(\|f^2\times p\|_\infty+f^2(\boldsymbol{k}))}{NA^{2D}}. \tag{52}$$

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

$$S_2 = \frac{\sigma^2R^{2D}}{NA^{2D}}\mathbb{E}\left[\prod_{j=1}^{D}\phi^2\left(\frac{\sin(R(k_j-X_{\cdot j}))}{R(k_j-X_{\cdot j})}\right)\right] \leq \frac{c'''\sigma^2R^D\|p\|_\infty R^D}{NA^{2D}}, \tag{53}$$

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

$$T_1 \leq C\left(\frac{(\|f^2 \times p\|_\infty + f^2(\boldsymbol{k}) + \sigma^2\|p\|_\infty)R^D}{N} + \frac{1}{R^{2(m+1)}}\right), \tag{54}$$

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

$$\begin{aligned}\text{var}(f_{N,R}(\boldsymbol{k})) \leq \frac{C_1'}{p^2(\boldsymbol{k})}\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]\left(\frac{1}{R^{2(m+1)}} + \frac{R^D\log(NR)}{N}\right)\\ + \frac{C_2'}{p^2(\boldsymbol{k})}\left(\frac{(f(\boldsymbol{k}) + C_3')R^D}{N} + \frac{1}{R^{2(m+1)}}\right),\end{aligned} \tag{55}$$

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:

$$\begin{aligned}\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right] \leq \frac{2(c')^2}{p^2(\boldsymbol{k})R^{2(m+1)}} + \frac{2\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]}{p^2(\boldsymbol{k})}\left(\frac{C_1}{R^{2(m+1)}} + \frac{C_2R^D}{N}\right)\\ + \frac{C_1'}{p^2(\boldsymbol{k})}\mathbb{E}\left[(f_{N,R}(\boldsymbol{k}) - f(\boldsymbol{k}))^2\right]\left(\frac{1}{R^{2(m+1)}} + \frac{R^D\log(NR)}{N}\right)\\ + \frac{C_2'}{p^2(\boldsymbol{k})}\left(\frac{(f(\boldsymbol{k}) + C_3')R^D}{N} + \frac{1}{R^{2(m+1)}}\right).\end{aligned}$$

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

D.4 Proof of Lemma 1

Invoking triangle inequality, we obtain that

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

If we denote $\mathbf{v}_i = \frac{R^D}{A^D}\prod_{j=1}^D \phi\left(\frac{\sin(R(k_j - k_{ij})}{R(k_j - k_{ij})}\right)$ 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 $|\mathbf{v}_i| \leq C^D R^D/A^D$ 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 $\text{var}(\mathbf{v}_i) \leq C'R^D$ where $C' > 0$ is some universal constant. Given these bounds of $\mathbf{v}_i$ and $\text{var}(\mathbf{v}_i)$, for any $t \in (0, C'']$ Bernstein's inequality shows that

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N \mathbf{v}_i - \mathbb{E}\left[\mathbf{v}_1\right]\right| \geq t\right) \leq 2\exp\left(-\frac{Nt^2}{2C'R^D + 2C^DR^Dt/(3A^D)}\right).$$

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

$$\mathbb{P}\left(\left|p^\phi_{N,R}(\boldsymbol{k}) - \mathbb{E}\left[p^\phi_{N,R}(\boldsymbol{k})\right]\right| \geq t\right) = \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N \mathbf{v}_i - \mathbb{E}\left[\mathbf{v}_1\right]\right| \geq t\right) \leq \delta. \tag{57}$$

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

$$\left|\mathbb{E}\left[p^\phi_{N,R}(\boldsymbol{k})\right] - p(\boldsymbol{k})\right| \leq c/R^{m+1}. \tag{58}$$

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 ϕ

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 ϕ such as $\phi(x) = |x|$, ReLU(x), and sigmoid(x) yield worse results.

Method	Valid PPL	Test PPL
Baseline dot-product (small)	33.15	34.29
FourierFormer, $\phi(x) = x^2$ (small)	32.09	33.10
FourierFormer, $\phi(x) = x^4$ (small)	31.86	32.85
FourierFormer, $\phi(x) = x^6$ (small)	**31.84**	**32.81**
FourierFormer, $\phi(x) = x$ (small)	not converge	not converge
FourierFormer, $\phi(x) = x^3$ (small)	not converge	not converge
FourierFormer, $\phi(x) = \|x\|$ (small)	33.12	34.18
FourierFormer, $\phi(x) =$ ReLU(x) (small)	33.87	35.01
FourierFormer, $\phi(x) =$ sigmoid(x) (small)	not converge	not converge

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.

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

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 ϕ 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.

Model	Runtime (Train) (miliseconds/sample)	GPU Memory (Train) (GB)	Runtime (Test) (miliseconds/sample)	GPU Memory (Test) (GB)
Baseline softmax (small)	5.41	1.43	1.53	0.94
FourierFormer (small)	6.00	1.43	1.70	0.94

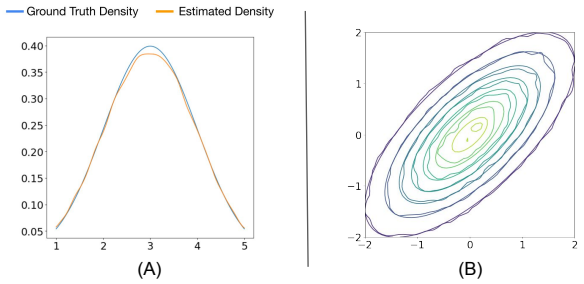

Figure 1. (A) 1-D and (B) 2-D Gaussian distributions and their estimated densities via Fourier Integral theorem.

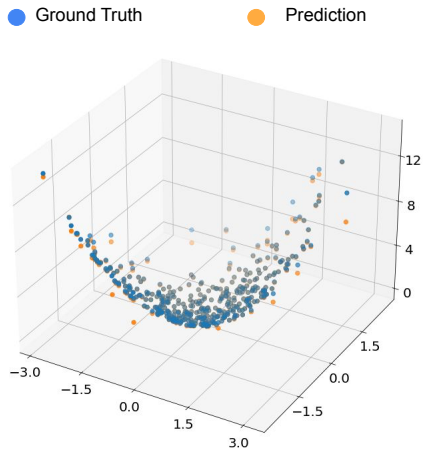

Figure 2: Non-parametric regression via the Fourier Integral theorem.

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 1.29×10^{-5} and 2.42×10^{-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.