Quantization Algorithms for Random Fourier Features: Supplemental Materials

In Section A, we provide details on the derivation of LM-RFF, Lloyd's algorithm for the construction and the exact quantizers as the output. In Section B, we present more analytical figures and results from Section 4 of the main paper. In Section C, we provide more details of the numerical experiments. Finally, all missing proofs are included in Section D.

A. Lloyd-Max (LM) Quantization: Derivation and Properties

We provide a detailed derivation of Lloyd-Max (LM) quantization scheme and its properties, which would be useful to our analysis. Recall that our proposed LM-RFF quantizers minimize the distortion defined as

$$D_Q = \int_{\mathcal{S}} (Q(z) - z)^2 f(z) dz,$$

where $f(z)$ is the signal distribution. Also, our b-bit fixed quantizer Q has borders $t_0 < ... < t_M$ and reconstruction levels $\mu_1 < ... < \mu_M$, with $M = 2^b$. Since the sine and cosine function are bounded within $[-1, 1]$, we have $t_0 = -1$ and $t_M = 1$. Thus the distortion is

$$D_Q = \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} (z - \mu_i)^2 f(z) dz.$$

Lloyd's algorithm finds a stationary point of above system. By setting the derivative of D_Q w.r.t. μ_i to 0

$$\frac{\partial D_Q}{\partial \mu_i} = -2 \int_{t_{i-1}}^{t_i} (z - \mu_i) f(z) dz = 0,$$

we obtain

$$\mu_i = \frac{\int_{t_{i-1}}^{t_i} z f(z) dz}{\int_{t_{t-1}}^{t_i} f(z) dz}.$$

We do the same thing for t_i (i.e., setting $\frac{\partial D_Q}{\partial t_i} = 0$) and get

$$t_i = \frac{\mu_i + \mu_{i+1}}{2}.$$

The following two useful properties hold for LM quantizers.

Property 1. $\mathbb{E}[z] = \mathbb{E}[Q(z)]$.

Property 2. $\mathbb{E}[Q(z)z] = \mathbb{E}[Q(z)^2]$.

Proof. For Property 1, we have

$$\begin{aligned} \mathbb{E}[Q(z)] &= \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \frac{\int_{t_{i-1}}^{t_i} z f(z) dx}{\int_{t_{i-1}}^{t_i} f(z) dz} f(z) dz \\ &= \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^{t_i} z f(z) dz = \mathbb{E}[z]. \end{aligned} \tag{9}$$

For Property 2, similarly we have

$$\begin{aligned}\mathbb{E}[Q(z)z] &= \sum_{i=1}^{M}\int_{t_{i-1}}^{t_i}\frac{\int_{t_{i-1}}^{t_i} zf(z)dx}{\int_{t_{i-1}}^{t_i} f(z)dz} zf(z)dz \\ &= \sum_{i=1}^{M}\int_{t_{i-1}}^{t_i}\frac{(\int_{t_{i-1}}^{t_i} zf(z)dx)^2}{(\int_{t_{i-1}}^{t_i} f(z)dz)^2} f(z)dz = \mathbb{E}[Q(z)^2].\end{aligned}$$

□

A.1. Details of LM-RFF Quantizer Construction

For completeness, we summarize our derivations introduced previously as the concrete steps for constructing LM-RFF in Algorithm 1. In our implementation, the algorithm terminates when the total absolute change in borders and reconstruction levels in two consecutive iterations is smaller than 10^{-5}. This convergence threshold can be set arbitrarily. In most cases, the running time of LM-RFF construction should be negligibly small.

For practitioners to use LM-RFF straightforwardly, we present the output LM-RFF quantizers with $b = 1, 2, 3, 4$ in Table 2.

Algorithm 1 Construction of LM-RFF quantizer

Input: Density $f_Z(z)$ (Theorem 2.1, (4)), number of bits b
Output: LM-RFF quantizer $\boldsymbol{t} = [t_0, ..., t_{2^b}]$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_{2^b}]$

Fix $t_0 = -1, t_{2^b} = 1$
While *true*
  For $i = 1$ to 2^b
    Update μ_i by $\mu_i = \frac{\int_{t_{i-1}}^{t_i} zf_Z(z)dz}{\int_{t_{i-1}}^{t_i} f_Z(z)dz}$
  End For
  For $i = 1$ to $2^b - 1$
    Update t_i by $t_i = \frac{\mu_i + \mu_{i+1}}{2}$
  End For
Until Convergence

Table 2. Constructed borders and reconstruction levels of LM-RFF quantizers, $b = 1, 2, 3, 4$, keeping three decimal places. Since the quantizers are symmetric about 0, we only present the positive part for conciseness.

b	**Borders**	**Reconstruction Levels**
1	$\{0, 1\}$	$\{0.637\}$
2	$\{0, 0.576, 1\}$	$\{0.297, 0.854\}$
3	$\{0, 0.286, 0.563, 0.819, 1\}$	$\{0.144, 0.428, 0.699, 0.939\}$
4	$\{0, 0.142, 0.283, 0.421, 0.557, 0.687, 0.811, 0.922, 1\}$	$\{0.071, 0.213, 0.353, 0.49, 0.624, 0.751, 0.87, 0.974\}$

B. More Analytical Figures in Section 4

B.1. More Figures on Mean and Variance of LM-RFF

In Figure 12, we present more figures on the bias of LM quantized estimators, corresponding to Theorem 4.2, Theorem 4.3. Same as in the main paper, we see that the proposed surrogates (Observations 4.1 and 4.2) align well with true biases. As b increases, the bias vanishes towards 0.

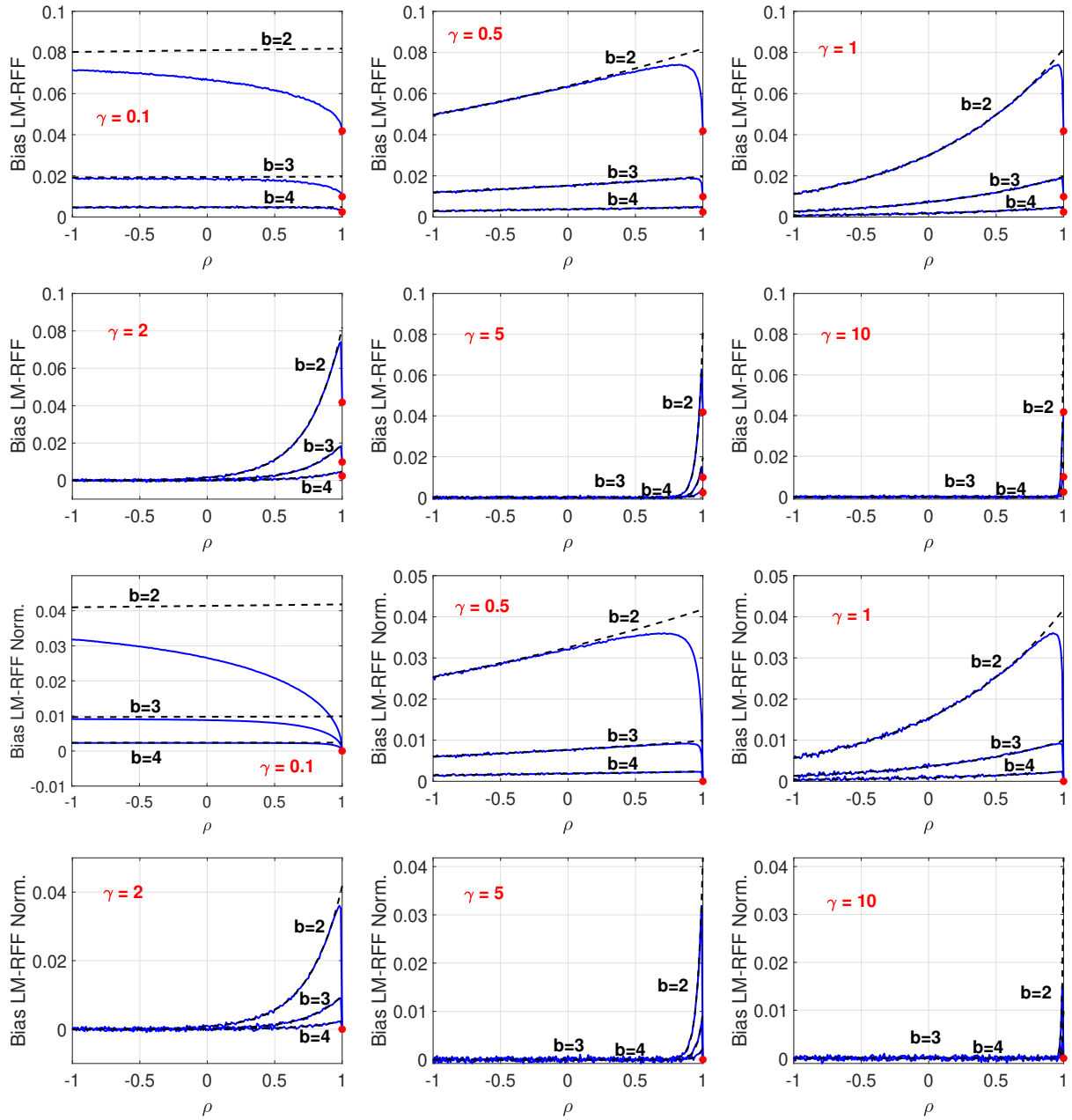

Figure 12. Observation 4.1 and Observation 4.2 (black dash curves) vs. empirical bias (blue curves) of LM-RFF. Red dots are the biases given in the theorems at specific ρ values.

In Figure 13, we provide more plots on variance of proposed LM-RFF estimators at more γ levels. As we expect, the variances of LM-RFF quantized estimators converge to the corresponding full-precision estimators as the number of bits b increases, i.e., $Var[\hat{K}_Q] \to Var[\hat{K}]$, $Var[\hat{K}_{n,Q}] \to Var[\hat{K}_n]$, as $b \to \infty$.

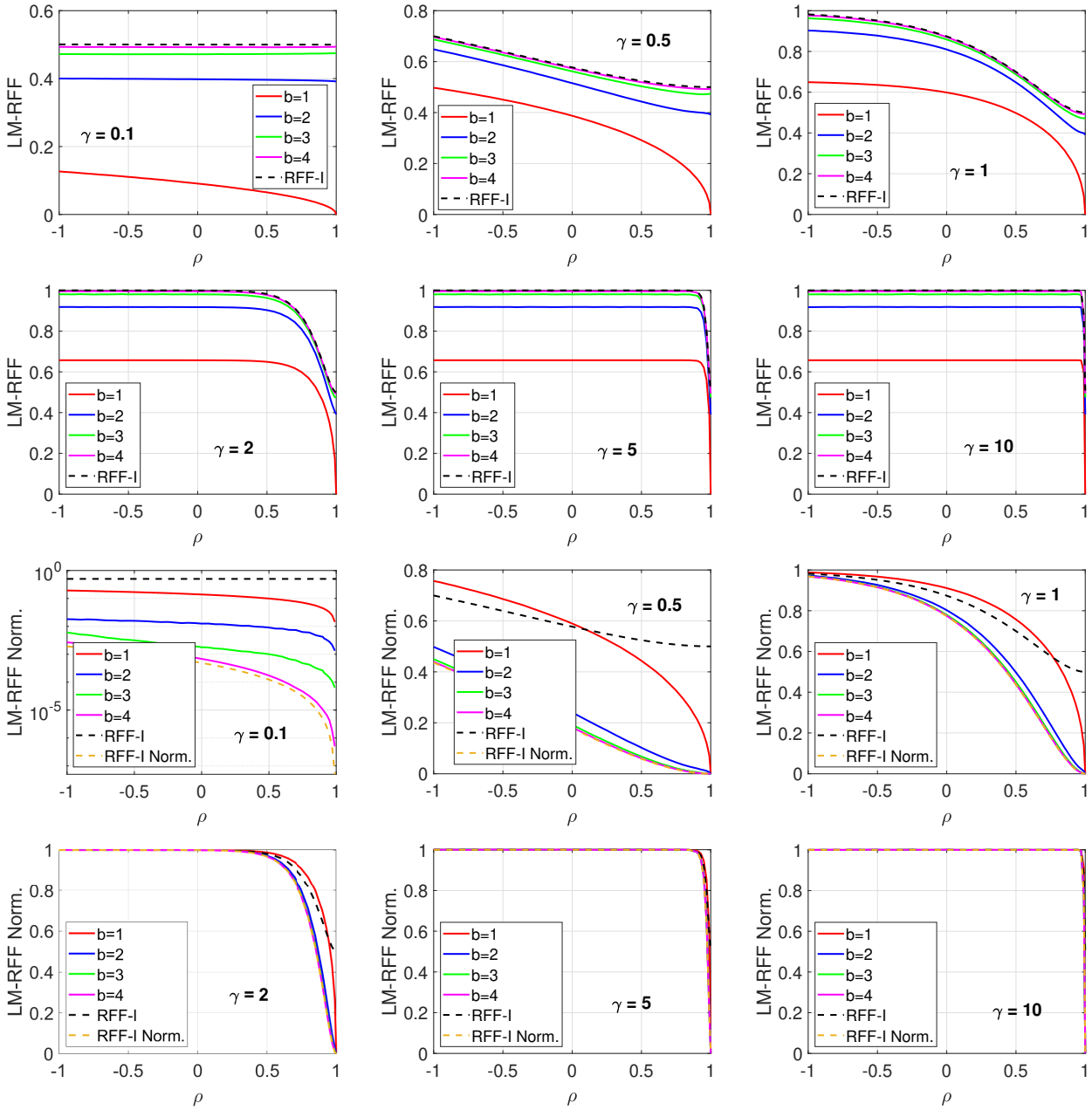

Figure 13. Variance (scaled by m) of LM-RFF and LM-RFF Norm. estimators with different γ and bits b. The dashed curves are the variances of full-precision counterparts.

B.2. Additional Example on the Monotonicity of LM-RFF Mean Estimation

In Lemma 4.5, we provide the concrete formula for computing the derivatives of $\mathbb{E}[g_1(z_x)g_2(z_y)]$ w.r.t. ρ, where g_1 and g_2 are two continuous functions, and z_x and z_y are the RFFs. In Theorem 4.6, we extend the result to discrete functions including LM-RFF quantizers. In Figure 14, we provide an additional example that validates Lemma 4.5 and Theorem 4.6, by approximating discrete LM-RFF quantization functions using continuous functions. Let $Q(x)$ be the 1-bit LM-RFF quantizer. We use continuous function $\tilde{g}(x) = \mu_2 \cdot sign(x)(1 - e^{-50|x|})$ as the surrogate to compute the "derivative" of Q. Here $\mu_2 = 0.697$ from Table 2 is the reconstruction level of 1-bit LM-RFF quantizer. We can draw similar conclusion as Figure 6 in the main paper: the derivative w.r.t. ρ is non-negative, and the theory matches the truth. Recall that in Theorem 4.6, with larger γ, monotonicity is guaranteed for larger ρ (Remark 4.1). In Figure 14, we see consistent pattern: for example, for $\gamma = 0.5$, we observe clearly the positive derivative on $\rho \in [0, 1]$, while for $\gamma = 5$, the derivative is almost zero until ρ is around 0.8.

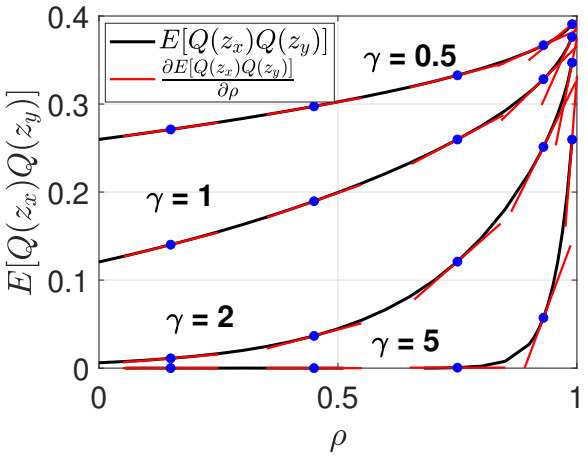

Figure 14. Validation of Lemma 4.5 and Theorem 4.6 at multiple γ. Black curves are the function value (expectation), and red lines are the theoretical derivatives. Here Q is the 1-bit LM-RFF quantizer. The derivative is approximated by the continuous approximation $\tilde{g}(x) = \mu_2 sign(x)(1 - e^{-50|x|})$, where μ_2 is the positive reconstruction level of Q.

C. More Experimental Details

In this section, we provide more implementation details of our empirical study.

Dataset description. For kernel SVM, `BASEHOCK` and `PCMAC` provided by ASU database (Li et al., 2016) are two subsets from the 20 NewsGroup dataset which are binary text datasets, each containing samples from 2 classes out of 20 categories. For both datasets, we process the samples by instance normalization. For kernel logistic regression, we use two popular datasets from LIBSVM library (Chang and Lin, 2011). The `Webspam` dataset, a benchmark dataset for spam detection, contains 175,000 training and test samples each, classified into "spam" and "not spam", where some negative samples are manually created by traversing some normal websites such as news articles. Unigram representation is adopted with 254 dimensions. Each sample is normalized to unit norm. The `CoverType` dataset predicts forest cover type from cartographic variables. We make a random 50/50 split to get the training and test set. We diretly train on the raw data without normalization. According to Theorem 2.1, our LM-RFF quantizer would work as well in this more general setting where data instances have arbitrary l_2 norms, which is justified by our numerical results.

Model training. We formally define the objective functions, and the tuned regularization parameters, of the learning problems in our experiments. We present linear models since our low-precision training is applied to linear learners. Suppose we have a dataset $\{\boldsymbol{u}_i \in \mathbb{R}^d, y_i\}_1^n$. For binary SVM classifier, assume $y_i \in \{-1, +1\}$ are the labels. Linear SVM solves the optimization problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{u}_i + b)),$$

where C is the penalization parameter. In logistic regression (with l_2 penalty), the problem is to solve

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n} \log(1 + e^{-y_i\boldsymbol{u_i}^T\boldsymbol{\beta}}) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2.$$

For ridge regression, the response y_i is real-valued. We minimize the squared error with l_2 regularization,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{i=1}^{n} (y_i - \boldsymbol{u_i}^T\boldsymbol{\beta})^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2,$$

where λ is a hyper-parameter controlling the penalization. All above three loss functions are convex, so Stochastic Gradient Descent (SGD) with appropriate stepsize finds the global minimizer. In linear learning, we train the models with input as the original data. In our approximated kernel learning setting, we solve the above three problems with the input vectors $\boldsymbol{u}_i$ as the random Fourier features (or the corresponding quantized RFFs).

The parameter γ for the Gaussian kernel is tuned over a fine grid from 0.001 to 100. For SVM experiments, the tuning parameter C for linear SVM is searched over a fine grid from 0.001 to 1000. For ridge regression, the penalization parameter λ is selected in the log-scale from $\{0, 1\text{e-}6, 1\text{e-}5, 1\text{e-}4, 1\text{e-}3, 1\text{e-}2, 1\text{e-}1\}$. We also tune the learning rate over a fine grid from 0.00001 to 0.1. These tuning procedures are applied to every single run using different compression method, b and m. For SVM experiments, we implement the popular LIBLINEAR (Chang and Lin, 2011) toolkit in MATLAB 2019a software. For KLR and KRR, we run SGD using PyTorch since it has well designed computing architecture for gradient-based optimization methods. For KLR runs we train for at least 50 epochs, and for KRR we train at least 100 epochs. The training is terminated after the test accuracy becomes stable (i.e., no more improvement for many epochs).

D. Proofs

D.1. Theorem 2.1

The following Lemma is a result of the convolution of normal and uniform distributions.

Lemma D.1. *Suppose $X \sim N(0,1)$ and $\tau \sim uniform(0, 2\pi)$ are independent, $\gamma > 0$. Then*

$$\gamma X + \tau \sim \frac{1}{2\pi}\left[\Phi(\frac{2\pi - y}{\gamma}) - \Phi(-\frac{y}{\gamma})\right].$$

Proof. We have the convolution of uniform and Gaussian distribution as

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} P(b = u, \gamma X = y - u)du \\ &= \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{\sqrt{2\pi}\gamma} e^{-\frac{(y-u)^2}{2\gamma^2}} du = \frac{1}{2\pi}\left[\Phi(\frac{2\pi - y}{\gamma}) - \Phi(-\frac{y}{\gamma})\right]. \end{aligned}$$

□

Proof. (of Theorem 2.1) Denote $Y = \gamma X + \tau$. We have

$$\begin{aligned} P\Big(Z \le z\Big) &= \sum_{k=-\infty}^{\infty} P\Big(2k\pi + \cos^{-1} z \le Y \le 2(k+1)\pi - \cos^{-1} z\Big) \\ &= \sum_{k=-\infty}^{\infty} \int_{2k\pi + \cos^{-1} z}^{2(k+1)\pi - \cos^{-1} z} f_Y(y)dy, \end{aligned}$$

where $f(y)$ is given by Lemma D.1. Let the density of Z be g_Z, and denote $t^* = \cos^{-1} z$. It follows that

$$\begin{aligned} g_Z(z) &= \sum_{k=-\infty}^{\infty} \frac{1}{\sqrt{1-z^2}}\Big[f_Y(2(k+1)\pi - t^*) + f_Y(2k\pi + t^*)\Big] \\ &= \frac{1}{2\pi\sqrt{1-z^2}} \sum_{k=-\infty}^{\infty} \underbrace{\Big[\Phi(\frac{t^* - 2k\pi}{\gamma}) - \Phi(\frac{t^* - 2(k+1)\pi}{\gamma}) + \Phi(\frac{-t^* - 2(k-1)\pi}{\gamma}) - \Phi(\frac{-t^* - 2k\pi}{\gamma})\Big]}_{\alpha_k} \\ &= \frac{1}{\pi\sqrt{1-z^2}}. \end{aligned} \tag{10}$$

To prove the last line, denote the term in the bracket as α_k. By cancellation, for any k_1, k_2, we have

$$\sum_{k=k_1}^{k_2} \alpha_k = \Big[\Phi(\frac{t^* - 2k_1\pi}{\gamma}) + \Phi(\frac{-t^* - 2(k_1-1)\pi}{\gamma}) - \Phi(\frac{t^* - 2(k_2+1)\pi}{\gamma}) - \Phi(\frac{-t^* - 2k_2\pi}{\gamma})\Big],$$

which equals to 2 in the limit as $k_1 \to -\infty, k_2 \to \infty$. Using a similar approach, we can show that Eq. (10) is exactly the density of the cosine of a uniform random variable on $[0, 2\pi]$. □

D.2. Theorem 2.2

We will use the following Lemma analogue to Lemma D.1 on the joint distribution.

Lemma D.2. *Denote $z_x = \gamma X + \tau$, $z_y = \gamma Y + \tau$ with $(X, Y) \sim N\big(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\big)$, $\tau \sim uniform(0, 2\pi)$. We have the joint distribution*

$$f(z_x, z_y) = \frac{1}{2\pi}\phi_{\sqrt{2(1-\rho)}\gamma}(z_x - z_y)\Big[\Phi(\frac{4\pi - (z_x + z_y)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(-\frac{z_x + z_y}{\gamma\sqrt{2(1+\rho)}})\Big].$$

Proof. Similar to the proof of Lemma D.1, we have

$$
\begin{aligned}
f(t_x, t_y) &= \frac{1}{2\pi}\int_0^{2\pi} P(\gamma x = t_x - u, \gamma y = t_y - u)du \\
&= \frac{1}{4\pi^2\gamma^2\sqrt{1-\rho^2}}\int_0^{2\pi} e^{-\frac{(t_x-u)^2-2\rho(t_x-u)(t_y-u)+(t_y-u)^2}{2(1-\rho^2)\gamma^2}}du \\
&= \frac{1}{4\pi^2\gamma^2\sqrt{1-\rho^2}}\int_0^{2\pi} e^{-\frac{2(1-\rho)(u^2-u(t_x+t_y))+t_x^2+t_y^2-2\rho t_x t_y}{2(1-\rho^2)\gamma^2}}du \\
&= \frac{1}{4\pi^2\gamma^2\sqrt{1-\rho^2}}\int_0^{2\pi} e^{-\frac{2(1-\rho)(u-\frac{t_x+t_y}{2})^2+\frac{1+\rho}{2}(t_x-t_y)^2}{2(1-\rho^2)\gamma^2}}du \\
&= \frac{1}{4\pi^2\gamma^2\sqrt{1-\rho^2}}e^{-\frac{(t_x-t_y)^2}{4(1-\rho)\gamma^2}}\int_0^{2\pi} e^{-\frac{(u-\frac{t_x+t_y}{2})^2}{(1+\rho)\gamma^2}}du \\
&= \frac{1}{2\pi}\phi_{\sqrt{2(1-\rho)}\gamma}(t_x-t_y)\Big[\Phi(\frac{4\pi-(t_x+t_y)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(-\frac{t_x+t_y}{\gamma\sqrt{2(1+\rho)}})\Big],
\end{aligned}
$$

where $\phi_{\sqrt{2(1-\rho)}\gamma}$ is the density of $N(0, 2(1-\rho)\gamma^2)$. □

Proof. (of Theorem 2.2) We will first prove the cosine function, and then prove similar result for applying sine functions, which will be useful for all subsequent proofs. Denote $Z_x = \cos(t_x), Z_y = \cos(t_y)$. Let $a_x^* = \cos^{-1}(z_x), a_y^* = \cos^{-1}(z_y)$. Denote $\phi = \phi_{\sqrt{2(1-\rho)}\gamma}$ for simplicity. We have

$$
P(Z_x \le z_x, Z_y \le z_y) = \sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\int_{2k_x\pi+a_x^*}^{2(k_x+1)\pi-a_x^*}\int_{2k_y\pi+a_y^*}^{2(k_y+1)\pi-a_y^*} f(t_x,t_y)dt_x dt_y.
$$

By Lemma D.2, it follows that

$$
\begin{aligned}
&f(z_x, z_y) \\
&= \frac{1}{2\pi}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\int_{2k_x\pi+a_x^*}^{2(k_x+1)\pi-a_x^*}\frac{1}{\sqrt{1-z_y^2}}\Bigg\{\phi(t_x-2(k_y+1)\pi+a_y^*)\Big[\Phi(\frac{4\pi-(t_x+2(k_y+1)\pi-a_y^*)}{\gamma\sqrt{2(1+\rho)}}) \\
&\quad - \Phi(-\frac{t_x+2(k_y+1)\pi-a_y^*}{\gamma\sqrt{2(1+\rho)}})\Big] + \phi(t_x-2k_y\pi-a_y^*)\Big[\Phi(\frac{4\pi-(t_x+2k_y\pi+a_y^*)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(-\frac{t_x+2k_y\pi+a_y^*}{\gamma\sqrt{2(1+\rho)}})\Big]\Bigg\}dt_x \\
&= \frac{1}{2\pi\sqrt{1-z_x^2}\sqrt{1-z_y^2}}\sum_{k_x}\sum_{k_y}\Bigg\{\phi(-a_x^*+a_y^*+2(k_x-k_y)\pi)\Big[\Phi(\frac{a_x^*+a_y^*-2(k_x+k_y)\pi)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(\frac{a_x^*+a_y^*-2(k_x+k_y+2)\pi}{\gamma\sqrt{2(1+\rho)}})\Big] \\
&\qquad + \phi(-a_x^*-a_y^*+2(k_x-k_y+1)\pi)\Big[\Phi(\frac{a_x^*-a_y^*-2(k_x+k_y-1)\pi)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(\frac{a_x^*-a_y^*-2(k_x+k_y+1)\pi}{\gamma\sqrt{2(1+\rho)}})\Big] \\
&\qquad + \phi(a_x^*+a_y^*+2(k_x-k_y-1)\pi)\Big[\Phi(\frac{-a_x^*+a_y^*-2(k_x+k_y-1)\pi)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(\frac{-a_x^*+a_y^*-2(k_x+k_y+1)\pi}{\gamma\sqrt{2(1+\rho)}})\Big] \\
&\qquad + \phi(a_x^*-a_y^*+2(k_x-k_y)\pi)\Big[\Phi(\frac{-a_x^*-a_y^*-2(k_x+k_y-2)\pi)}{\gamma\sqrt{2(1+\rho)}}) - \Phi(\frac{-a_x^*-a_y^*-2(k_x+k_y)\pi}{\gamma\sqrt{2(1+\rho)}})\Big]\Bigg\} \\
&\overset{(a)}{=} \frac{1}{2\pi\sqrt{1-z_x^2}\sqrt{1-z_y^2}}\sum_{k=-\infty}^{\infty}\Big[\phi(-a_x^*+a_y^*+2k\pi)+\phi(-a_x^*-a_y^*+2k\pi)+\phi(a_x^*+a_y^*+2k\pi)+\phi(a_x^*-a_y^*+2k\pi)\Big] \\
&= \frac{1}{\pi\sqrt{1-z_x^2}\sqrt{1-z_y^2}}\sum_{k=-\infty}^{\infty}\Big[\phi(a_x^*-a_y^*+2k\pi)+\phi(a_x^*+a_y^*+2k\pi)\Big],
\end{aligned}
$$

where (a) is derived by writing the summations $\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\{\cdot\}$ into $\sum_{l=-\infty}^{\infty}\sum_{k_x=-\infty}^{\infty}\{\cdot\}$ with $l = k_x - k_y$ and canceling terms, along with the symmetry of $\phi(\cdot)$. This gives the joint density of z_x and z_y.

For the sine counterpart, with some abuse of notation, let us denote $z_x = \sin(t_x)$ and $z_y = \sin(t_y)$ from now on. Using similar argument, we have

$$P(Z_x \leq z_x, Z_y \leq z_y) = \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \int_{(2k_x+1)\pi-\sin^{-1}(z_x)}^{2(k_x+1)\pi+\sin^{-1}(z_x)} \int_{(2k_y+1)\pi-\sin^{-1}(z_y)}^{2(k_y+1)\pi+\sin^{-1}(z_y)} f(t_x, t_y) dt_x dt_y.$$

After simplification, we finally arrive at

$$f(z_x, z_y) = \frac{1}{\pi\sqrt{1-z_x^2}\sqrt{1-z_y^2}} \sum_{k=-\infty}^{\infty} \Big[\phi(\sin^{-1}(z_x) - \sin^{-1}(z_y) + 2k\pi) + \phi(\sin^{-1}(z_x) + \sin^{-1}(z_y) + (2k+1)\pi)\Big]. \tag{11}$$

Considering $Z_x = \sin(t_x), Z_y = \sin(t_y)$. Since $\sin^{-1}(x) = \frac{\pi}{2} - \cos^{-1}(x)$, we can substitute into the density to derive

$$\begin{aligned} f(z_x, z_y) &= \frac{1}{\pi\sqrt{1-z_x^2}\sqrt{1-z_y^2}} \sum_{k=-\infty}^{\infty} \Big[\phi(\cos^{-1}(z_x) - \cos^{-1}(z_y) + 2k\pi) + \phi(\cos^{-1}(z_x) + \cos^{-1}(z_y) + (2k+2)\pi)\Big] \\ &== \frac{1}{\pi\sqrt{1-z_x^2}\sqrt{1-z_y^2}} \sum_{k=-\infty}^{\infty} \Big[\phi(a_x^* - a_y^* + 2k\pi) + \phi(a_x^* + a_y^* + 2k\pi)\Big], \end{aligned}$$

which is the same as the previous cosine transformation. This completes the proof. □

D.3. Proposition 2.3

Proof. Let us denote $\sigma = \sqrt{2(1-\rho)}\gamma$ for simplicity. By symmetry and exchangeability of f, to prove the desired result, it suffices to consider the case where both z_x and z_y are positive, i.e., $(z_x, z_y) \in (0, 1]^2$. Define the notation $a_x^* = \sin^{-1}(z_x) \geq 0, a_y^* = \sin^{-1}(z_y) \geq 0$. From (11), we deduct

$$\begin{aligned} & f(z_x, z_y) - f(z_x, -z_y) \\ & \propto \sum_{k=-\infty}^{\infty} \Big[\phi_\sigma(a_x^* - a_y^* + 2k\pi) + \phi_\sigma(a_x^* + a_y^* + (2k+1)\pi) \\ & \qquad - \phi_\sigma(a_x^* + a_y^* + 2k\pi) - \phi_\sigma(a_x^* - a_y^* + (2k+1)\pi)\Big] \\ & = \sum_{k=0}^{\infty} (-1)^k \Big[\phi_\sigma(k\pi + d) - \phi_\sigma(k\pi + s) + \phi_\sigma((k+1)\pi - s) - \phi_\sigma((k+1)\pi - d)\Big], \\ & = \phi_\sigma(d) - \phi_\sigma(s) + \sum_{k=1}^{\infty} \Big[\phi_\sigma(k\pi - s) - \phi_\sigma(k\pi - d) - \phi_\sigma(k\pi + d) + \phi_\sigma(k\pi + s)\Big], \\ & \triangleq \phi_\sigma(d) - \phi_\sigma(s) + \sum_{k=1}^{\infty} M_k, \end{aligned} \tag{12}$$

where we let $d = a_x^* - a_y^*$ and $d = a_x^* + a_y^*$, and we use the fact that $\phi_\sigma(-x) = \phi_\sigma(x)$. Note that, we consider $z_y > 0$ so that $d \neq s$, since when $z_y = 0$ we trivially have $f(z_x, 0) = f(z_x, 0)$. For now, we assume that $z_x \geq z_y > 0$, such that d and s are defined on the domain $0 < s \leq \pi$ and $0 \leq d < \min\{s, \pi - s\}$. Since

$$\phi'_\sigma(x) = -\frac{x}{\sqrt{2\pi}\sigma^3} e^{-\frac{x^2}{2\sigma^2}}, \quad \phi''_\sigma(x) = -\frac{x^2 - \sigma^2}{\sqrt{2\pi}\sigma^5} e^{-\frac{x^2}{2\sigma^2}},$$

we know that ϕ_σ is piecewise concave on $(0, \sigma)$ and piecewise convex on (σ, ∞). Thus,

$$\phi_\sigma(a) - \phi_\sigma(a+g) \geq \phi_\sigma(c) - \phi_\sigma(c+g) \tag{13}$$

for any $\sigma \le a \le c$ and $g \ge 0$. The equality holds only when $a = c$ or $g = 0$. Consequently, under the assumption that $\sigma \le \pi$, $M_k \ge 0$ for $k \ge 2$ since $2\pi - s \ge \sigma$, where the equality holds only when $d = s$, i.e., $z_y = 0$. Furthermore, the piecewise convexity of $\phi_\sigma(\cdot)$ and (13) imply that for $\sigma \le a < c$,

$$\frac{(c-a)c}{\sigma^2} e^{-\frac{c^2}{2\sigma^2}} < e^{-\frac{a^2}{2\sigma^2}} - e^{-\frac{c^2}{2\sigma^2}} < \frac{(c-a)a}{\sigma^2} e^{-\frac{a^2}{2\sigma^2}}. \tag{14}$$

Also note that the function e^{-x} is convex on the real line, which gives for $\forall a < c$,

$$\frac{(c-a)(c+a)}{2\sigma^2} e^{-\frac{c^2}{2\sigma^2}} < e^{-\frac{a^2}{2\sigma^2}} - e^{-\frac{c^2}{2\sigma^2}} < \frac{(c-a)(c+a)}{2\sigma^2} e^{-\frac{a^2}{2\sigma^2}}. \tag{15}$$

Now that $M_k > 0$ for $k \ge 2$, evaluating (12) we obtain

$$\begin{aligned}
(12) &> \phi_\sigma(d) - \phi_\sigma(s) + M_1 \\
&\overset{(a)}{>} \frac{1}{\sqrt{2\pi}\sigma^3}\Big[\frac{(s-d)(s+d)}{2} e^{-\frac{s^2}{2\sigma^2}} + \frac{(s-d)(2\pi-s-d)}{2} e^{-\frac{(\pi-d)^2}{2\sigma^2}} - (s-d)(\pi+d)e^{-\frac{(\pi+d)^2}{2\sigma^2}}\Big] \\
&\overset{(b)}{\ge} \frac{s-d}{\sqrt{2\pi}\sigma^3}\Big[\pi e^{-\frac{(\pi-d)^2}{2\sigma^2}} - (\pi+d)e^{-\frac{(\pi+d)^2}{2\sigma^2}}\Big],
\end{aligned}$$

where (a) uses (14) and (15), and (b) is because $s \le \pi - d$. It is easy to verify that the ratio

$$\Big(\pi e^{-\frac{(\pi-d)^2}{2\sigma^2}}\Big) \Big/ \Big((\pi+d)e^{-\frac{(\pi+d)^2}{2\sigma^2}}\Big) = \frac{\pi}{\pi+d} e^{\frac{2\pi d}{\sigma^2}} \ge 1$$

for $\sigma \le \pi$ and $0 \le d < \min\{s, \pi - s\} < \frac{\pi}{2}$. Therefore, we have proved that $f(z_x, z_y) > f(z_x, -z_y)$, for $z_x \ge z_y > 0$. Now, by exchangeability and symmetry of f, we have

$$f(z_y, z_x) = f(z_x, z_y) > f(z_x, -z_y) = f(-z_x, z_y) = f(z_y, -z_x).$$

Therefore, our result also holds for $z_y \ge z_x > 0$. The proof is now complete.

□

D.4. Theorem 4.1

Proof. Denote the StocQ quantizer as Q. For each RFF z, assume $z \in [t_{i-1}, t_i]$ for some i. We can then write $Q(z) = z + \epsilon$, where

$$\mathbb{E}[\epsilon] = t_i \frac{z - t_{i-1}}{t_i - t_{i-1}} + t_{i-1} \frac{t_i - z}{t_i - t_{i-1}} - z = 0.$$

Thus, it follows that

$$\begin{aligned}
Var[\epsilon] = \mathbb{E}[\epsilon^2] &= t_i^2 \frac{z - t_{i-1}}{t_i - t_{i-1}} + t_{i-1}^2 \frac{t_i - z}{t_i - t_{i-1}} - z^2 \\
&= (t_i - z)(z - t_{i-1}).
\end{aligned}$$

For two data vectors u, v, let $F^{StocQ}(u) = \sqrt{2}Q(z_u)$ and $F^{StocQ}(v) = \sqrt{2}Q(z_v)$, where $z_u = \cos(w^T u + \tau)$ and $z_v = \cos(w^T v + \tau)$ follows the distribution f given by Theorem 2.2. We can write $Q(z_u) = z_u + \epsilon_u$, $Q(z_v) = z_v + \epsilon_v$ where ϵ_u and ϵ_v are independent. Let $\hat{=} F^{StocQ}(u) F^{StocQ}(v)$. We have

$$\begin{aligned}
\mathbb{E}[\hat{K}_{StocQ}] &= 2\mathbb{E}[(z_u + \epsilon_u)(z_v + \epsilon_v)] \\
&= 2\mathbb{E}[z_u z_v] = K(u, v),
\end{aligned}$$

implying that StocQ estimate is unbiased. The variance factor can be computed as

$$\begin{aligned}
Var[\hat{K}_{StocQ}] &= 4\mathbb{E}[(z_u + \epsilon_u)^2 (z_v + \epsilon_v)^2] - K(u, v)^2 \\
&= 4\mathbb{E}[z_u^2 \epsilon_v^2 + z_v^2 \epsilon_u^2 + \epsilon_u^2 \epsilon_v^2] + Var[\hat{K}] \triangleq A + Var[\hat{K}],
\end{aligned} \tag{16}$$

where $Var[\hat{K}]$ is the variance of full-precision RFF kernel estimator. Obviously, $A > 0$, thus StocQ estimator always has larger variance than full-precision RFF. Continuing our analysis,

$$\begin{aligned}\mathbb{E}[z_u^2\epsilon_v^2] &= \mathbb{E}_{z_u,z_v} z_u^2 \mathbb{E}[\epsilon_v^2|z_v] \\ &= \int_{-1}^{1} dz_u \Big(\sum_{j=1}^{2^b-1} \int_{t_{j-1}}^{t_j} (t_j - z_v)(z_v - t_{j-1}) z_u^2 f(z_u, z_v) dz_v \Big) \\ &= \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i} \Big((t_{j-1}+t_j) z_v z_u^2 - z_v^2 z_u^2 - t_{j-1}t_j z_u^2 \Big) f(z_u, z_v) dz_u dz_v.\end{aligned}$$

By symmetry of density function f, we know that $\mathbb{E}[z_v^2\epsilon_u^2] = \mathbb{E}[z_u^2\epsilon_v^2]$. It remains to compute $\mathbb{E}[\epsilon_u^2\epsilon_v^2]$. By similar reasoning, we have

$$\begin{aligned}\mathbb{E}[\epsilon_u^2\epsilon_v^2] &= \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i} (t_i - z_u)(z_u - t_{i-1})(t_j - z_v)(z_v - t_{j-1}) f(z_u, z_v) dz_u dz_v \\ &= \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i} \Big((t_{i-1}+t_i)(t_{j-1}+t_j) z_u z_v - (t_{i-1}+t_i) z_u z_v^2 - (t_{j-1}+t_j) z_v z_u^2 + z_u^2 z_v^2 \\ &\quad - (t_{i-1}+t_i) t_{j-1} t_j z_u - (t_{j-1}+t_j) t_{i-1} t_i z_v + t_{j-1}t_j z_u^2 + t_{i-1}t_i z_v^2 + t_{i-1}t_i t_{j-1} t_j \Big) f(z_u, z_v) dz_u dz_v \\ &\overset{(a)}{=} \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i} \Big((t_{i-1}+t_i)(t_{j-1}+t_j) z_u z_v - 2(t_{j-1}+t_j) z_v z_u^2 + z_u^2 z_v^2 \\ &\quad + 2t_{j-1}t_j z_u^2 + t_{i-1}t_i t_{j-1} t_j \Big) f(z_u, z_v) dz_u dz_v,\end{aligned}$$

where equation (a) is due to the symmetry of density f and the borders $t_0 < ... < t_{2^b-1}$. Substituting above expressions into (16) and cancelling terms, we obtain

$$\begin{aligned}A &= 4 \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i} \Big((t_{i-1}+t_i)(t_{j-1}+t_j) z_u z_v + t_{i-1}t_i t_{j-1} t_j - z_u^2 z_v^2 \Big) f(z_u, z_v) dz_u dz_v \\ &= 4 \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \Big[(t_{i-1}+t_i)(t_{j-1}+t_j)\kappa_{i,j} + t_{i-1}t_i t_{j-1} t_j p_{i,j} \Big] - 4\mathbb{E}[z_u^2 z_v^2].\end{aligned}$$

Therefore,

$$Var[\hat{K}_{StocQ}] = 4 \sum_{i=1}^{2^b-1}\sum_{j=1}^{2^b-1} \Big[(t_{i-1}+t_i)(t_{j-1}+t_j)\kappa_{i,j} + t_{i-1}t_i t_{j-1} t_j p_{i,j} \Big] - K(u,v)^2.$$

The proof is completed by noting that StocQ estimator is the average of i.i.d. Bernoulli random variables. □

D.5. Theorem 4.2

Proof. We start by recalling some preliminaries on functional analysis. The *Chebyshev polynomials* (Borwein and Erdélyi, 1995) of the first kind are defined through trigonometric identities

$$T_n(\cos(x)) = \cos(n\cos(x)),$$

where admit the following recursion,

$$\begin{aligned}&T_0(x) = 1, \quad T_1(x) = x, \\ &T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x), \quad i \geq 2.\end{aligned}$$

$\{T_0, T_1, ...\}$ forms an orthogonal basis of the function space on $[-1, 1]$ with finite number of discontinuities. Precisely, define the inner product w.r.t. measure $\frac{1}{\sqrt{1-x^2}}$ as

$$\langle f(x), g(x)\rangle = \int_{-1}^{1} f(x)g(x)\frac{1}{\sqrt{1-x^2}}dx.$$

Then orthogonality holds:

$$\int_{-1}^{1} T_i(x)T_j(x)\frac{1}{\sqrt{1-x^2}}dx = \begin{cases} 0, & i \neq j, \\ \pi, & i = j = 0, \\ \frac{\pi}{2}, & i = j \neq 0. \end{cases}$$

By Chebyshev functional decomposition, our LM quantizer can be written as

$$Q(x) = \sum_{k=0}^{\infty} \alpha_k T_k(x),$$

where α_k are computed through the inner products,

$$\begin{aligned} \alpha_0 &= \frac{2}{\pi}\int_{-1}^{1} Q(x)T_0(x)\frac{dx}{\sqrt{1-x^2}} = 0, \\ \alpha_1 &= \frac{2}{\pi}\int_{-1}^{1} Q(x)T_1(x)\frac{dx}{\sqrt{1-x^2}} = 1 - 2D, \\ \alpha_2 &= \frac{2}{\pi}\int_{-1}^{1} Q(x)T_2(x)\frac{dx}{\sqrt{1-x^2}} = 0, \\ \alpha_3 &= \frac{2}{\pi}\int_{-1}^{1} Q(x)T_3(x)\frac{dx}{\sqrt{1-x^2}}, \\ &... \end{aligned}$$

with D the distortion of Q given in equation (7) of the main paper. Firstly, it is easy to show that $|\mathbb{E}[T_i(z_x)T_j(z_y)]| \leq \mathbb{E}[T_i(z_x)^2] = \frac{1}{2}$. Note that $\alpha_k = 0$ when k is even because $T_k(x)$ is even function and $Q(x)$ is odd. Recall u, v are two normalized data vectors with correlation ρ. Denote $z_x = \cos(\gamma x + \tau)$ and $z_y = \cos(\gamma y + \tau)$ with distribution $f(z_x, z_y)$, where $(x, y) \sim N\big(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\big)$, $\tau \sim uniform(0, 2\pi)$. It follows that

$$\begin{aligned} \mathbb{E}[\sqrt{2}Q(z_x)\sqrt{2}Q(z_y)] &= 2\int_{-1}^{1}\int_{-1}^{1} Q(z_x)Q(z_y)f(z_x, z_y)dz_x dz_y \\ &= 2\int_{-1}^{1}\int_{-1}^{1}(\sum_{i=1,odd}^{\infty}\alpha_i T_i(z_x))(\sum_{j=1,odd}^{\infty}\alpha_j T_j(z_y))f(z_x, z_y)dz_x dz_y \\ &= (1-2D)^2 K(u, v) + 2\sum_{i=1,odd}^{\infty}\sum_{j=3,odd}^{\infty}\alpha_i\alpha_j\int_{-1}^{1}\int_{-1}^{1} T_i(z_x)T_j(z_y)f(z_x, z_y)dz_x dz_y. \end{aligned} \tag{17}$$

This proves the first part. There is an intrinsic constraint on α_i, $i = 3, 5, ...$ First, we can compute the cosine of $Q(x)$ and each $T_i(x)$ as

$$\begin{aligned} c_i &= \frac{\int_{-1}^{1} Q(x)T_i(x)\frac{dx}{\sqrt{1-x^2}}}{\sqrt{\int_{-1}^{1} Q(x)^2\frac{dx}{\sqrt{1-x^2}}\int_{-1}^{1} T_i(x)^2\frac{dx}{\sqrt{1-x^2}}}} \\ &= \frac{\frac{\pi}{2}\alpha_i}{\sqrt{(\frac{1}{2}-D)\pi}\sqrt{\frac{\pi}{2}}} \\ &= \frac{\alpha_i}{\sqrt{1-2D}}. \end{aligned}$$

Since the Chebyshev polynomials form an orthogonal basis of function space on $[-1,1]$, it holds that $\sum_{i=0}^{\infty} c_i^2 = 1$. Therefore, we have $\sum_{i=0}^{\infty} \alpha_i^2 = 1-2D$. Now that $\alpha_i = 0$ when i is even, and $\alpha_1 = 1-2D$, we then have $\sum_{i=3,odd}^{\infty} \alpha_i^2 = 1-2D-(1-2D)^2 = 2D(1-2D)$.

When $\rho = 0$, from (17), it is easy to see that all the integrals would be zero by independence. Thus, the estimated kernel $\mathbb{E}[\sqrt{2}Q(z_x)\sqrt{2}Q(z_y)] = (1-2D)^2 K(u,v)$.

When $\rho = 1$ ($K(u,v) = 1$), we have $\int_{-1}^{1} T_i(z_x)T_j(z_x)f(z_x)dz_x = 0$ for $i \neq j$ by orthogonality of Chebyshev polynomials, where $f(z_x)$ is the marginal distribution of z_x. It follows that

$$\begin{aligned}\mathbb{E}[\sqrt{2}Q(z_x)\sqrt{2}Q(z_y)] &= (1-2D)^2 + \sum_{i=3,odd}^{\infty} \alpha_i^2 \\ &= (1-2D)^2 + 2D(1-2D) \\ &= 1-2D.\end{aligned}$$

This completes the proof of the theorem. □

D.6. Theorem 4.3

Proof. Denote $\boldsymbol{w} = \cos(\gamma\boldsymbol{x}+\tau)$, $\boldsymbol{z} = \cos(\gamma\boldsymbol{y}+\tau)$, with $(\boldsymbol{x},\boldsymbol{y})$ are random vectors with i.i.d. entries from $N(0, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix})$, and $\tau \sim uniform(0, 2\pi)$ is also a vector with i.i.d. entries. Recall the notation $\zeta_{s,t} = \mathbb{E}[Q(w_1)^s Q(z_1)^t]$, where Q is our LM-RFF quantizer. By Taylor expansion at the expectations, we have as $m \to \infty$,

$$\begin{aligned}\mathbb{E}[\hat{K}_{n,Q}] &= \frac{\mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} Q(w_i)Q(z_i)]}{\mathbb{E}[\sqrt{\frac{1}{m^2}\|Q(\boldsymbol{w})\|^2\|Q(\boldsymbol{z})\|^2}]} + \mathcal{O}(\frac{1}{m}) \\ &\triangleq \frac{\zeta_{1,1}}{\mathbb{E}[\sqrt{\Lambda}]} + \mathcal{O}(\frac{1}{m}).\end{aligned}$$

Applying Taylor expansion again,

$$\begin{aligned}\mathbb{E}[\sqrt{\Lambda}] &= \mathbb{E}\Big[\sqrt{\mathbb{E}[\Lambda]} + \frac{\Lambda - \mathbb{E}[\Lambda]}{2\sqrt{\mathbb{E}[\Lambda]}} + \mathcal{O}((\Lambda - \mathbb{E}[\Lambda])^2)\Big] \\ &= \mathbb{E}[\Lambda] + \mathcal{O}(\frac{1}{m}), \quad m \to \infty.\end{aligned}$$

Furthermore, we have the expectation of Λ as

$$\begin{aligned}\mathbb{E}[\Lambda] &= \frac{1}{m^2}\mathbb{E}\Big[\big(\sum_{i=1}^{m} Q(w_i)^2\big)\big(\sum_{i=1}^{m} Q(z_i)^2\big)\Big] \\ &= \frac{1}{m^2}\Big[\sum_{i\neq j} Q(w_i)^2 Q(z_j)^2 + \sum_{i=1}^{m} Q(w_i)^2 Q(z_i)^2\Big] \\ &= \frac{m-1}{m}\mathbb{E}[Q(w_1)^2 Q(z_2)^2] + \frac{1}{m}\mathbb{E}[Q(w_1)^2 Q(z_1)^2] \\ &= \zeta_{2,0}^2, \quad m \to \infty.\end{aligned}$$

Consequently, we obtain

$$\mathbb{E}[\hat{K}_{n,Q}] = \frac{\zeta_{1,1}}{\zeta_{2,0}}, \quad m \to \infty.$$

This completes the proof for asymptotic mean. With a little abuse of notation, let $\hat{K}_{n,Q} = \frac{a}{\sqrt{bc}}$, with

$$a = \frac{\langle Q(\boldsymbol{w}), Q(\boldsymbol{z})\rangle}{k},\ b = \frac{\|Q(\boldsymbol{w})\|^2}{k},\ c = \frac{\|Q(\boldsymbol{z})\|^2}{k}.$$

We have

$$\begin{aligned}
\mathbb{E}[a] = \zeta_{1,1}, \quad Var[a] &= \frac{\zeta_{2,2} - \zeta_{1,1}^2}{m}, \\
\mathbb{E}[b] = \zeta_{2,0} = \mathbb{E}[c], \quad Var[b] &= \frac{\zeta_{4,0} - \zeta_{2,0}^2}{m} = Var[c], \\
Cov(a,b) &= \frac{1}{m^2}\mathbb{E}[(\sum_{i=1}^{m} Q(w_i)Q(z_i))(\sum_{i=1}^{m} Q(w_i)^2)] - \zeta_{1,1}\zeta_{2,0} \\
&= \frac{m\zeta_{3,1} + m(m-1)\zeta_{1,1}\zeta_{2,0}}{m^2} - \zeta_{1,1}\zeta_{2,0} \\
&= \frac{\zeta_{3,1} - \zeta_{1,1}\zeta_{2,0}}{m} = Cov(a,c), \\
Cov(b,c) &= \frac{\zeta_{2,2} - \zeta_{2,0}^2}{m}.
\end{aligned}$$

We can formulate the covariance matrix

$$Cov(a,b,c) = \frac{1}{m}\begin{pmatrix} \zeta_{2,2} - \zeta_{1,1}^2 & \zeta_{3,1} - \zeta_{1,1}\zeta_{2,0} & \zeta_{3,1} - \zeta_{1,1}\zeta_{2,0} \\ \zeta_{3,1} - \zeta_{1,1}\zeta_{2,0} & \zeta_{4,0} - \zeta_{2,0}^2 & \zeta_{2,2} - \zeta_{2,0}^2 \\ \zeta_{3,1} - \zeta_{1,1}\zeta_{2,0} & \zeta_{2,2} - \zeta_{2,0}^2 & \zeta_{4,0} - \zeta_{2,0}^2 \end{pmatrix}.$$

The gradient vector at the expectations is

$$\nabla \hat{K}_{n,Q}(\mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[c]) = (\frac{1}{\zeta_{2,0}}, -\frac{\zeta_{1,1}}{2\zeta_{2,0}^2}, -\frac{\zeta_{1,1}}{2\zeta_{2,0}^2}).$$

By Taylor expansion, it holds that

$$Var[\hat{K}_{n,Q}] = \nabla \hat{K}_{n,Q}(\mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[c])^T Cov(a,b,c) \nabla \hat{K}_{n,Q}(\mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[c]) + \mathcal{O}(\frac{1}{m^2}).$$

The theorem is proved by plugging in the expressions. □

D.7. Theorem 4.4

Proof. Let $z_x = \cos(\gamma X + \tau)$, $z_y = \cos(\gamma Y + \tau)$ where $(X, Y) \sim N\big(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\big)$, $\tau \sim uniform(0, 2\pi)$. Denote $\zeta_{s,t} = \mathbb{E}[Q(z_x)^s Q(z_y)^t]$. Recalling Theorem 4.2 and Theorem 4.3, we have asymptotically (omitting lower order terms)

$$\begin{aligned}
\mathbb{E}[\hat{K}_Q] = 2\zeta_{1,1}, \quad Var[\hat{K}_Q] &= \frac{4}{m}(\zeta_{2,2} - \zeta_{1,1}^2), \\
\mathbb{E}[\hat{K}_{n,Q}] = \frac{\zeta_{1,1}}{\zeta_{2,0}}, \quad Var[\hat{K}_{n,Q}] &= \frac{1}{m}\Big(\frac{\zeta_{2,2}}{\zeta_{2,0}^2} - \frac{2\zeta_{1,1}\zeta_{3,1}}{\zeta_{2,0}^3} + \frac{\zeta_{1,1}^2(\zeta_{4,0} + \zeta_{2,2})}{2\zeta_{2,0}^4}\Big).
\end{aligned}$$

Thus, we can compute the debiased estimator variance as (after simplification)

$$\begin{aligned}
Var^{db}[\hat{K}_Q] &= \frac{K(u,v)^2}{m}\Big(\frac{\zeta_{2,2}}{\zeta_{1,1}^2} - 1\Big), \\
Var^{db}[\hat{K}_{n,Q}] &= \frac{K(u,v)^2}{m}\Big(\frac{\zeta_{2,2}}{\zeta_{1,1}^2} - \frac{2\zeta_{3,1}}{\zeta_{1,1}\zeta_{2,0}} + \frac{\zeta_{4,0}\zeta_{2,2}}{2\zeta_{2,0}^2}\Big).
\end{aligned}$$

Taking the difference, we obtain

$$\begin{aligned}
Var^{db}[\hat{K}_Q] - Var^{db}[\hat{K}_{n,Q}] &\propto 4\zeta_{2,0}\zeta_{3,1} + \zeta_{1,1}(\zeta_{4,0} + \zeta_{2,2}) - 2\zeta_{1,1}\zeta_{2,0}^2 \\
&\geq 4\zeta_{2,0}\zeta_{3,1} + \zeta_{1,1}(\zeta_{2,2} - \zeta_{2,0}^2) \triangleq M(\rho),
\end{aligned}$$

where the inequality is due to the fact that $\zeta_{4,0} - \zeta_{2,0}^2 = Var[Q^2(z_x)] \geq 0$. Here we denote M as a function of ρ. At $\rho = 0$, we have

$$\zeta_{3,1} = 0, \quad \zeta_{2,2} = \zeta_{2,0}^2,$$

so that $M(0) = 0$. At $\rho = 1$, it holds that

$$\zeta_{3,1} = \zeta_{2,2} = \zeta_{4,0},$$

hence $M(1) > 0$. Notice that $Q(\cdot)$ and $Q^3(\cdot)$ are non-decreasing odd functions, and $Q^2(\cdot)$ is a even function. For $\rho \in [0, 1]$, since $\sqrt{2(1-\rho)}\gamma \leq \sqrt{2}\gamma \leq \pi$ by assumption, it follows from Theorem 4.6 that $\zeta_{1,1}$, $\zeta_{2,2}$ and $\zeta_{3,1}$ are all increasing in ρ on $[0, 1]$. Consequently, $M(\rho) > 0$ for any $\rho \in [0, 1]$. The desired result thus follows. □

D.8. Lemma 4.5

Lemma D.3 (Stein's Lemma)**.** *Suppose $X \sim N(\mu, \sigma^2)$, and g is a differentiable function such that $\mathbb{E}[g(X)(X-\mu)]$ and $\mathbb{E}[g'(X)]$ exist. Then, $\mathbb{E}[g(X)(X-\mu)] = \sigma^2\mathbb{E}[g'(X)]$.*

Proof. (of Lemma 4.5) We use the technique of Gaussian interpolation and Stein's Lemma. First, we formulate $Y = \gamma\rho X + \gamma\sqrt{1-\rho^2}Z$ where $Z \sim N(0, 1)$ independent of X. By continuity and boundedness of g_1 and g_2, it holds that

$$\begin{aligned}
&\frac{\partial \mathbb{E}[g_1(\cos(s_x))g_2(\cos(s_y))]}{\partial \rho} \\
&= \frac{\partial \mathbb{E}_{X,Z,\tau}[g_1(\cos(\gamma X + \tau))g_2(\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau))]}{\partial \rho} \\
&= -\mathbb{E}_{X,Z,\tau}\Big[\underbrace{g_1(\cos(\gamma X + \tau))g_2'(\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau))\sin(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau)}_{\Upsilon(X,Z;\rho)}\big(\gamma X - \frac{\gamma\rho Z}{\sqrt{1-\rho^2}}\big)\Big].
\end{aligned}$$

We analyze two parts respectively. By Lemma D.3 and law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}_{X,Z,\tau}[\Upsilon(X, Z; \rho)\gamma X] \\
&= \mathbb{E}_{Z,\tau}\mathbb{E}_X\Big[-\gamma^2 g_1'(\cos(\gamma X + \tau))\sin(\gamma X + \tau)g_2'(\cos(\gamma Y + \tau))\sin(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau) \\
&\quad - \gamma^2\rho g_1(\cos(\gamma X + \tau))g_2''(\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau))\sin^2(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau) \\
&\quad + \gamma^2\rho g_1(\cos(\gamma X + \tau))g_2'(\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau))\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau)|Z, \tau\Big], \qquad (18)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{X,Z,\tau}\big[\Upsilon(X, Z; \rho)\frac{\gamma\rho Z}{\sqrt{1-\rho^2}}\big] \\
&= \mathbb{E}_{X,\tau}\mathbb{E}_Z\Big[-\gamma^2\rho g_1(\cos(\gamma X + \tau))g_2''(\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau))\sin^2(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau) \\
&\quad + \gamma^2\rho g_1(\cos(\gamma X + \tau))g_2'(\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau))\cos(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau)|X, \tau\Big]. \qquad (19)
\end{aligned}$$

Combining (18) and (19), we get

$$\begin{aligned}
&\frac{\partial \mathbb{E}[g_1(\cos(s_x))g_2(\cos(s_y))]}{\partial \rho} \\
&= \mathbb{E}_{X,Z,\tau}\Big[\gamma^2 g_1'(\cos(\gamma X + \tau))\sin(\gamma X + \tau)g_2'(\cos(\gamma Y + \tau))\sin(\gamma\rho X + \gamma\sqrt{1-\rho^2}Z + \tau)\Big] \\
&= \gamma^2\mathbb{E}_{X,Y,\tau}\big[g_1'(\cos(s_x))\sin(s_x)g_2'(\cos(s_y))\sin(s_y)\big],
\end{aligned}$$

which gives the desired expression.

To prove the monotonicity, suppose that g_1 and g_2 are increasing odd or non-constant even functions. So, $g_1'(-x)g_2'(-x) = g_1'(x)g_2'(x) > 0, \forall x \in [-1, 1]$. Assume $\sqrt{2(1-\rho)}\gamma \leq \pi$, and denote $f(x, y)$ as the joint density given by Theorem 1. We can write

$$\begin{aligned}
\frac{\partial \mathbb{E}[g_1(\cos(s_x))g_2(\cos(s_y))]}{\partial \rho} &= \gamma^2 \int_{-1}^{1}\int_{-1}^{1} z_x z_y g_1'(\sqrt{1-z_x^2}) g_2'(\sqrt{1-z_y^2}) f(z_x, z_y) dz_x dz_y \\
&\overset{(a)}{=} 2\gamma^2 \Big(\int_{0}^{1}\int_{0}^{1} z_x z_y g_1'(\sqrt{1-z_x^2}) g_2'(\sqrt{1-z_y^2}) f(z_x, z_y) dz_x dz_y \\
&\qquad + \int_{0}^{1}\int_{-1}^{0} z_x z_y g_1'(\sqrt{1-z_x^2}) g_2'(\sqrt{1-z_y^2}) f(z_x, z_y) dz_x dz_y \Big) \\
&= 2\gamma^2 \int_{0}^{1}\int_{0}^{1} z_x z_y g_1'(\sqrt{1-z_x^2}) g_2'(\sqrt{1-z_y^2}) [f(z_x, z_y) - f(z_x, -z_y)] dz_x dz_y \\
&\overset{(b)}{>} 0,
\end{aligned}$$

where (a) is due to the symmetry of f and g, and (b) is a consequence of Proposition 2.3 that $f(z_x, z_y) > f(z_x, -z_y)$ for all $z_x, z_y \in (0, 1]^2$, provided that $\sqrt{2(1-\rho)}\gamma \leq \pi$. The proof is complete.

□

D.9. Theorem 4.6

Proof. Since Q_1 and Q_2 both are non-decreasing and have finite number of discontinuities, by Baire's Characterization Theorem (Baire and Denjoy, 1905; Hausdorff, 1991), we know that each of them is the point-wise limit of a sequence of continuous increasing functions. Suppose that $\{g_{1,n}\}$ and $\{g_{2,n}\}$ are two sequences of continuous increasing functions such that as $n \to \infty$, $g_{1,n} \to Q_1$ and $g_{2,n} \to Q_2$ with point-wise convergence. By dominated convergence theorem, we have

$$\begin{aligned}
\frac{\partial \mathbb{E}[Q_1(z_x)Q_2(z_y)]}{\partial \rho} &= \frac{\partial \mathbb{E}[\lim_{n\to\infty} g_{1,n}(z_x) \lim_{n\to\infty} g_{2,n}(z_y)]}{\partial \rho} \\
&= \lim_{n\to\infty} \frac{\partial \mathbb{E}[g_{1,n}(z_x)g_{2,n}(z_y)]}{\partial \rho} > 0,
\end{aligned}$$

where Lemma 4.5 is adopted for continuous $g_{1,n}$ and $g_{2,n}$ functions. □