DETERMINANTAL POINT PROCESSES ON SPHERES: MULTIVARIATE LINEAR STATISTICS

RENJIE FENG[1], FRIEDRICH GÖTZE[2], AND DONG YAO

In memory of Steve Zelditch (1953-2022)

Abstract. In this paper, we will derive the first and 2nd order Wiener chaos decomposition for the multivariate linear statistics of the determinantal point processes associated with the spectral projection kernels on the unit spheres S^d. We will first get a graphical representation for the cumulants of multivariate linear statistics for any determinantal point process. The main results then follow from the very precise estimates and identities regarding the spectral projection kernels and the symmetry of the spheres.

1. Introduction

The determinantal point process is an important class of point processes with applications in random matrix theory, statistical mechanics, quantum mechanics, etc. It's also called the Slater determinant in quantum mechanics that is to describe the wave function of a multi-fermionic system. In this paper, we will consider determinantal point processes on the unit spheres associated with the spectral projection kernels of the Laplace operator with respect to the standard round metric. Such spectral projection kernels can be represented in terms of the spherical harmonics, which are one of the most fundamental wave functions in quantum mechanics to describe particles confined to the spheres.

Let Φ be a point process sampled on the space $\mathcal{X}$. The k-th joint intensity function ρ_k of the point process Φ is defined by

$$\mathbb{E}\Big[\sum_{(x_1,\dots,x_k)\in\Phi_*^k} f(x_1,\dots,x_k)\Big] = \int_{\mathcal{X}^k} f(x_1,\dots,x_k)\rho_k(x_1,\dots,x_k)dx_1\dots dx_k, \quad (1)$$

where f is any bounded measurable function and the set

$$\Phi_*^k := \{(x_1,\dots,x_k) : x_i \in \Phi, \forall 1 \le i \ne j \le k, x_i \ne x_j\}. \quad (2)$$

If Φ is a determinantal point process associated with some kernel function K, then its k-th joint intensity function reads

$$\rho_k(x_1,\dots,x_k) = \det\Big(K(x_i,x_j)_{1\le i\le j\le k}\Big), \quad (3)$$

where $K(x_i,x_j)_{1\le i,j\le k}$ is a $k\times k$ matrix whose (i,j) entry is $K(x_i,x_j)$.

In this paper we will focus on the case when K is the spectral projection kernel on the unit sphere S^d with $d\ge 2$, defined as follows. The Laplace operator on S^d with respect to the standard round metric has discrete spectrum $\Big\{\lambda_n = -n(n+d-1), n =$

1) Research supported by DFG GO 420/12-1.
2) Research supported by SFB 1283/2 2021 – 317210226.

$0, 1, 2, ...\Big\}$. Here, the round metric is the pullback of the Euclidean metric under the inclusion map $i : S^d \to \mathbb{R}^{d+1}$. For a given eigenvalue λ_n, the corresponding eigenfunctions are called the spherical harmonics of level n. Let $\mathcal{H}_n(S^d)$ be the space of the spherical harmonics of level n. Then one has [2]

$$k_n := \dim \mathcal{H}_n(S^d) = \frac{2n+d-1}{n+d-1}\binom{n+d-1}{d-1}, \tag{4}$$

which admits the asymptotic estimate (by $d \geq 2$)

$$k_n \sim 2n^{d-1}/\Gamma(d). \tag{5}$$

Let K_n be the spectral projection

$$K_n : L^2(S^d) \to \mathcal{H}_n(S^d), \tag{6}$$

and we denote by $K_n(x, y)$ its kernel.

Now we define a determinantal point process Φ_n on S^d associated with the kernel $K_n(x, y)$. Here the total number of points in Φ_n is almost surely k_n. Note that Φ_n can be alternatively defined by sampling k_n points on S^d according to the probability density

$$\frac{1}{k_n!} \det\Big(K_n(x_i, x_j)_{1\leq i,j\leq k_n}\Big). \tag{7}$$

Given a function $f(x_1, .., x_k)$ of $k \geq 1$ variables, we define the multivariate linear statistics

$$L_n f := \sum_{(x_1,\ldots,x_k)\in\Phi^k_{n,*}} f(x_1, \ldots, x_k), \tag{8}$$

where

$$\Phi^k_{n,*} := \{(x_1, \ldots, x_k) : x_i \in \Phi_n,\, x_i \neq x_j,\, \forall 1 \leq i \neq j \leq k\}. \tag{9}$$

Multivariate linear statistics of this form are usually called U-statistics.

For $1 \leq i \leq k$, we define the *i-margin* function f_i by integrating f with respect to all variables over S^d except x_i, i.e.,

$$f_i(x) = \int_{(S^d)^{k-1}} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_k)dx_1 \cdots dx_{i-1}dx_{i+1} \cdots dx_k. \tag{10}$$

Here, we denote by dx the volume element with respect to the standard round metric on S^d. If $k = 1$, the 1-margin function is defined to be f itself.

For $1 \leq i < j \leq k$, we define the (i, j)-*margin* function $f_{i,j}$ to be

$$f_{i,j} = \int_{(S^d)^{k-2}} f(x_1, \ldots, x_k)dx_1 \cdots dx_{i-1}dx_{i+1} \cdots dx_{j-1}dx_{j+1} \cdots dx_k. \tag{11}$$

In this article, we will study the limiting distribution of the multivariate linear statistics $L_n f$. We first give an asymptotic expansion for the expectation of $L_n f$.

Theorem 1. *Let $f(x_1, .., x_k)$ be a bounded function of k variables. We have*

$$\begin{aligned}\mathbb{E}(L_n f) = &\left(\frac{k_n}{s_d}\right)^k \int_{(S^d)^k} f(x_1, \ldots, x_k)dx_1 \cdots dx_k \\ &- \frac{k_n^{k-1}}{s_d^k} \sum_{1\leq i<j\leq k} \frac{2^{d-1}}{\Gamma(d)\pi}\Gamma\Big(\frac{d}{2}\Big)^2 \int_{S^d}\int_{S^d} \frac{f_{i,j}(x, y)}{\sin^{d-1}(\arccos(x \cdot y))}dxdy \\ &+ o(n^{(d-1)(k-1)}),\end{aligned} \tag{12}$$

where $s_d = 2\pi^{\frac{d+1}{2}}/\Gamma(\frac{d+1}{2})$ *is the surface area of* S^d.

By estimating the growth order of the cumulants of $L_n f$, we can prove the following central limit theorem for $L_n f$.

Theorem 2. *Let* f *be a bounded function of* k *variables on* S^d. *Assume that*

$$F(x) := \sum_{i=1}^{k} f_i(x) \tag{13}$$

is not constant almost everywhere in $x \in S^d$, *then it holds that*

$$\lim_{n\to\infty} \frac{1}{k_n^{2k-1}} \mathrm{Var}(L_n f) = \frac{2^{d-2}}{s_d^{2k}\Gamma(d)\pi} \Gamma\Big(\frac{d}{2}\Big)^2 \int_{S^d}\int_{S^d} \frac{(F(x)-F(y))^2}{\sin^{d-1}(\arccos(x\cdot y))} dxdy > 0. \tag{14}$$

In addition, $L_n f$ *is asymptotically normal, i.e.,*

$$\frac{L_n f - \mathbb{E}(L_n f)}{(\mathrm{Var}(L_n f))^{\frac{1}{2}}} \xrightarrow{d} N(0,1), \tag{15}$$

where $N(0,1)$ *is the standard Gaussian distribution and the notation* $\xrightarrow{d}$ *means the convergence in distribution.*

Combining Theorem 1 and Theorem 2, we have the following corollary.

Corollary 1. *Under the assumption of Theorem 2,*

$$\left(L_n f - \left(\frac{k_n}{s_d}\right)^k \int_{(S^d)^k} f(x_1,\dots,x_k) dx_1 \cdots dx_k \right) \mathrm{Var}(L_n f)^{-1/2} \xrightarrow{d} N(0,1).$$

When the assumption of Theorem 2 fails, i.e., $F(x)$ is constant almost everywhere, the right hand side of (14) will be degenerate, i.e., $\mathrm{Var}(L_n f)$ will have strictly smaller growth order than $k_n^{2k-1} = \Theta(n^{(d-1)(2k-1)})$. For such degenerate case, our next theorem shows that for a class of test functions, the limiting distribution is given by a mixture of centered chi-square distributions, i.e., the 2nd order Wiener Chaos.

We now consider the following two invariance conditions on the bounded test function $f(x_1,\dots,x_k)$, $k \geq 2$.

- f is invariant under permutations, i.e.,
$$f(x_1,\dots,x_k) = f(x_{\sigma(1)},\dots,x_{\sigma(k)}), \forall\, \sigma \in \mathrm{Sym}(k). \tag{16}$$
- We assume that the $(1,2)$-margin function $f_{1,2}(x_1,x_2)$ only depends on their spherical distance $\mathrm{dist}(x_1,x_2)$ (abbreviated as $\mathrm{d}(x_1,x_2)$), i.e.,
$$f_{1,2}(x_1,x_2) = f_{1,2}(x_1',x_2'),\ \ \forall\, \mathrm{d}(x_1,x_2) = \mathrm{d}(x_1',x_2'). \tag{17}$$

We will show that if the test function f satisfies these two assumptions, then $F(x)$ must be constant on the sphere, and thus the variance will be degenerate.

As a remark, the condition (16) is not an essential one. We can always symmetrize a function f by considering the average

$$\bar{f}(x_1,\dots,x_k) = \frac{1}{k!} \sum_{\sigma\in\mathrm{Sym}(k)} f(x_{\sigma(1)},\dots,x_{\sigma(k)}),$$

and this will yield $L_n f = L_n \bar{f}$ by (8).

There is an important class of test functions that satisfy these two assumptions. For example, given $\delta > 0$, if we choose

$$f(x_1, x_2) = \mathbf{1}[\mathrm{d}(x_1, x_2) < \delta], \tag{18}$$

where the indicator function is equal to 1 if the distance $\mathrm{d}(x_1, x_2) < \delta$ and 0 otherwise, then the random variable $L_n f$ will be the number of pairs of random points whose distances are less than δ. Similarly, if we take

$$f(x_1, x_2, x_3) = \mathbf{1}[\mathrm{d}(x_1, x_2) < \delta,\, \mathrm{d}(x_1, x_3) < \delta,\, \mathrm{d}(x_2, x_3) < \delta], \tag{19}$$

then $L_n f$ will count the number of triangles where the three vertices of the triangle are within distance δ. These types of counting statistics are useful tools to study the topology of random complexes built over random point processes, due to its connections with Betti numbers, e.g., [4, 6, 11]. Our main result Theorem 3 below implies that such types of counting statistics of the determinantal point process on S^d converge to the 2nd order Wiener chaos.

Under conditions (16) and (17) we can determine the growth order of $\mathrm{Var}(L_n f)$ and find the limiting distribution of $L_n f$. We define the function

$$\widehat{h}(x, y) := \int_{S^d} (f_{1,2}(x, y) - f_{1,2}(x, z)) \sin^{-(d-1)}(\arccos(z \cdot y)) dz. \tag{20}$$

We will see that $\widehat{h}$ is a bounded symmetric function, and thus we can consider it as a Hilbert-Schmidt integral operator acting on $L^2(S^d)$. Then this operator is compact and self-adjoint. Therefore we have the spectral decomposition

$$\widehat{h}(x, y) = \sum_{j=1}^{\infty} z_j w_j(x) w_j(y), \tag{21}$$

where $\{z_j, j \geq 1\}$ are eigenvalues of the operator, and $\{w_j, j \geq 1\}$ are the corresponding eigenfunctions which form an orthonormal basis of $L^2(S^d)$.

The following theorem states that the multivariate linear statistics will tend to a mixture of centered chi-squared distributions in the degenerate case.

Theorem 3. *For any bounded function $f(x_1, \ldots, x_k)$ with $k \geq 2$ satisfying conditions* (16) *and* (17)*, we have*

$$\lim_{n\to\infty} \frac{\mathrm{Var}(L_n f)}{k_n^{2k-2}} = \frac{2C_d^2 k^2 (k-1)^2}{\Gamma(d)^2 s_d^{2k}} \int_{(S^d)^2} \widehat{h}(x, y)^2 dx dy, \tag{22}$$

where the constant $C_d := \frac{2^{d-2}\Gamma(d/2)^2}{\pi}$. Furthermore, we have

$$\Big(L_n f - \mathbb{E}(L_n f)\Big) \left(\frac{k_n}{s_d}\right)^{-k} \left(\frac{C_d k(k-1)}{n^{d-1}}\right)^{-1} \xrightarrow{d} \sum_{i=1}^{\infty} z_i(\chi_i - 1)/2, \tag{23}$$

where $\chi_i, i \geq 1$ are independent chi-squared random variables with one degree of freedom and $\sum_{i=1}^{\infty} z_i(\chi_i - 1)/2$ is understood as the L^2-limit of $\sum_{i=1}^{N} z_i(\chi_i - 1)/2$ as $N \to \infty$.

Similar to Corollary 1, using the fact that $k_n \sim 2n^{d-1}/\Gamma(d)$, and Theorems 1 and 3, we deduce the following result.

Corollary 2. *Under the assumptions of Theorem 3, we have*

$$\left(\frac{k_n}{s_d}\right)^{-k}\left(\frac{C_d k(k-1)}{n^{d-1}}\right)^{-1}\left(L_n f-\left(\frac{k_n}{s_d}\right)^k\int_{(S^d)^k} f(x_1,\ldots,x_k)dx_1\cdots dx_k\right.$$
$$\left.+\frac{k_n^{k-1}}{s_d^k}\sum_{1\le i<j\le k}\frac{2^{d-1}}{\Gamma(d)\pi}\Gamma\Big(\frac{d}{2}\Big)^2\int_{S^d}\int_{S^d}\frac{f_{i,j}(x,y)}{\sin^{d-1}(\arccos(x\cdot y))}dxdy\right) \tag{24}$$
$$\xrightarrow{d}\sum_{i=1}^{\infty} z_i(\chi_i-1)/2.$$

Note that the limiting distribution can be rewritten in the form of the 2nd order Wiener chaos

$$\sum_{i=1}^{\infty} z_i H_2(X_i)/2, \tag{25}$$

where $H_2(x)=x^2-1$ is the Hermite polynomial of degree 2, and X_i are independent and identically distributed (i.i.d.) standard Gaussian random variables $N(0,1)$.

There is a vast literature on the univariate linear statistics of determinantal point processes, e.g., [7, 8, 9]. There are also very few works that give conditions for a Gaussian limit of multivariate linear statistics, e.g., [3]. But to the best of our knowledge, Theorem 3 is the very first result on the multivariate linear statistics for determinantal point processes beyond the Gaussian limit case.

Theorem 2 and Theorem 3 are proved by the method of cumulants. We will first derive a graphical representation for the cumulants of the multivariate linear statistics for any determinantal point process in Lemma 1, which generalizes the well-known formula for the univariate case (see (42) below). This graphical representation allows us to study the asymptotic properties of the cumulants by the off-diagonal decay of the spectral projection kernel, where we have to prove Lemma 3 and Lemma 4 bounding multiple integrals over the product of kernels. Exact identities and asymptotic expansions of the spectral projection kernels combined with the symmetry of the underlying space of the sphere are two crucial ingredients for our proofs. For example, we repeatedly use the facts that the spectral projection kernel is constant on the diagonal, and it satisfies very precise off-diagonal estimates for all length scales, e.g., (53); the important fact that the integral operator $\widehat{h}(x,y)$ defined in (20) is symmetric is partially due to the symmetry of the sphere, etc.

Contrary to the i.i.d. point process, the determinantal point process has the negative association property. But our main results Theorem 2 and Theorem 3 are still analogs of the classical Wiener chaos decomposition in the theory of U-statistics for i.i.d random variables.

Given i.i.d. random variables $X_1,\cdots,X_n$, Hoeffding's form for U-statistics is the following (normalized) multivariate linear statistics,

$$U_n^k(g)=\binom{n}{k}^{-1}\sum_{1\le i_1<\cdots<i_k\le n} g(X_{i_1},\ldots,X_{i_k}),$$

where g is a symmetric real-valued function of k variables.

Without loss of generality, we assume $\mathbb{E}(g(X_1,..,X_k))=0$. Then Hoeffding in 1948 proved that, if the variance $\mathrm{Var}(g(X_1,..,X_k))<\infty$, then the following central

limit theorem holds (Corollary 11.5 in [5]),

$$n^{1/2}U_n^k(g) \xrightarrow{\mathrm{d}} N(0, k^2\delta_1). \tag{26}$$

Here, the constant δ_1 is the variance

$$\delta_1 = \mathrm{Var}(g_1(X_1)),$$

where

$$g_1(x) := \mathbb{E}(g(x, X_2, .., X_k)).$$

If the variance δ_1 vanishes, that is the limit of U-statistics for i.i.d. random variables is degenerate, then a χ^2-limit theorem holds for the rescaled statistics. To be more precise, we suppose that $g_1(x) = \mathbb{E}g(x, X_2, .., X_k) = 0$ and $\mathbb{E}g^2(X_1, .., X_k) < \infty$, then we have (Corollary 11.5 in [5]),

$$nU_n^k(g) \xrightarrow{\mathrm{d}} \binom{k}{2} \sum_{i=1}^{\infty} \lambda_i H_2(Y_i), \tag{27}$$

where $H_2(x) = x^2 - 1$ is the Hermite polynomials of degree 2, Y_i are i.i.d. standard Gaussian random variables, and λ_i are eigenvalues of the integral operator A defined as follows. Let $d\mu$ be the probability density of the random variable X_1 and set

$$g_2(x, y) := \mathbb{E}g(x, y, X_3, .., X_k).$$

For any bounded measurable function f, the operator A is define by

$$(Af)(y) = \int g_2(x, y)f(x)d\mu(x). \tag{28}$$

The formats of results (26) and (27) are almost identical to Theorem 2 and Theorem 3, respectively. The roles of $g_1(x_1)$ and $g_2(x_1, x_2)$ are replaced by the i-margin function $f_i(x)$ and the (i, j)-margin function $f_{i,j}(x, y)$ respectively; when the variance vanishes, both the limiting distributions are the linear eigenvalue combination of $H_2(Y_i)$, where the role of the symmetric integral operator A is replaced by $\widehat{h}(x, y)$.

In general, U_n^k may exhibit the convergence in distribution to the Wiener chaos with arbitrary order (Theorem 11.3 in [5]). For example, for the primitive completely degenerate case where

$$g(x_1, \ldots, x_k) = \prod_{i=1}^{k} \mathfrak{g}(x_i)$$

with $\mathbb{E}\mathfrak{g}(X_1) = 0$ and $\mathbb{E}\mathfrak{g}^2(X_1) = \sigma^2 < \infty$, one has the convergence

$$\frac{n^{k/2}U_n^k(g)}{\sigma^k} \xrightarrow{\mathrm{d}} H_k(Y), \tag{29}$$

where $H_k(x)$ is the Hermite polynomial of degree k and Y is the standard Gaussian random variable.

Therefore, we may expect that the multivariate linear statistics of the determinantal point process associated with the spectral projection kernel on S^d also admits some kind of Wiener chaos decomposition. Actually, our method, especially the representation formula in Lemma 1, can be applied to any other determinantal point process such as CUE, GUE, the complex Ginibre ensemble in random matrix theory and Gaussian analytic functions in random polynomial theory. And the similar results may hold as well, but note that one has to change the conditions

especially (17) for the test functions to others according to the symmetry and the invariance of the underlying space and the kernel.

Notation. In this paper, we use C (or c) to denote some constants independent of n, whose specific values may change from line to line. For a sequence of numbers a_n and b_n, we write $a_n = o(b_n)$ if $b_n \neq 0$ and $\lim_{n\to\infty} a_n/b_n = 0$; $a_n = O(b_n)$ if there exists some constant C such that $|a_n| \leq C\,|b_n|$; $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$; $a_n \sim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$.

2. A graphical representation of cumulants

In this section we will derive a graphical representation of the cumulants for the multivariate linear statistics of any determinantal point process.

Given a random variable X, its m-th cumulant $Q_m(X)$ is defined to be the coefficient in the formal expansion of $\log \mathbb{E}\exp(\mathrm{i}tX)$,

$$\log \mathbb{E}\exp(\mathrm{i}tX) = \sum_{m=1}^{\infty} \frac{Q_m(X)}{m!}(\mathrm{i}t)^m. \tag{30}$$

A *partition* of a set S is an unordered collection $R = \{R_1, \ldots, R_\ell\}$ of nonempty subsets of S where ℓ is some positive integer not exceeding $|S|$. In addition, R satisfies the following two conditions:

- $R_i \cap R_j = \emptyset$ for $i \neq j$.
- $\cup_{i=1}^{\ell} R_i = S$.

Let m be any positive integer. We denote by $\Pi(m)$ the set of partitions of $\{1, 2, \cdots, m\}$. The moments of X can be derived from its cumulants as follows,

$$\mathbb{E}(X^m) = \sum_{R=\{R_1,\ldots,R_\ell\}\in\Pi(m)} Q_{|R_1|}\cdots Q_{|R_\ell|}. \tag{31}$$

On the other hand, the cumulants can be expressed by moments as

$$Q_m(X) = \sum_{R=\{R_1,\ldots,R_\ell\}\in\Pi(m)} (-1)^{\ell-1}(\ell-1)!\Pi_{i=1}^{\ell}\mathbb{E}X^{|R_i|}. \tag{32}$$

Some simple properties of cumulants include

$$Q_1(X) = \mathbb{E}(X),\ Q_2(X) = \mathrm{Var}(X),\ Q_m(cX) = c^m Q_m(X).$$

If X is a Gaussian random variable, then $Q_m(X) = 0$ for all $m \geq 3$.

Similarly to the method of moments, to show that X_n converges in distribution to X, it suffices to prove that the m-th cumulant of X_n converges to $Q_m(X)$ for all fixed m (as long as the limit is uniquely determined by its cumulants). For the special case that X is Gaussian distributed and X_n has mean 0, it suffices to prove

$$\lim_{n\to\infty} \frac{Q_m(X_n)}{\mathrm{Var}(X_n)^{\frac{m}{2}}} = 0$$

for all sufficiently large m ([9, Lemma 3]).

Let Φ be a determinantal point process on the space $\mathcal{X}$ associated with the kernel $K(x, y)$. In the followings, we will derive a formula for the cumulants of the multivariate linear statistics. We will expand the m-th power of the multivariate linear statistics and express it in the form of (31), then the formula for the cumulants can be found directly from this expression.

To expand $(\sum_{(x_1,\ldots,x_k)\in\Phi_*^k} f(x_1,..,x_k))^m$, we have km points $x_1,\ldots,x_{mk}$ (counting multiplicities) appearing in the product $f(x_1,\ldots,x_k)\cdots f(x_{mk-k+1},\ldots,x_{mk})$. We write $y_{i,j} := x_{(i-1)k+j}$ for $1\le i\le m$, $1\le j\le k$, and set $\mathbf{y}_i := (y_{i,1},\ldots,y_{i,k})$. Then we have

$$\left(\sum_{(x_1,\ldots,x_k)\in\Phi_*^k} f(x_1,\ldots,x_k)\right)^m = \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_m\in\Phi_*^k} f(\mathbf{y}_1)\cdots f(\mathbf{y}_m). \tag{33}$$

We first introduce a notation: given any positive integer p, we define the set

$$[p] := \big\{1,\ldots,p\big\}.$$

To find the relations among the points $x_1,\ldots,x_{mk}$, we define by

$$M(m,k) := \mathrm{Map}([m],[km]_*^k)$$

the set of all maps from $[m]$ to

$$[km]_*^k := \Big\{(i_1,\ldots,i_k)\in[km]^k : i_j\neq i_\ell,\ \forall\, 1\le j<\ell\le k\Big\}. \tag{34}$$

To be more precise, let $\mathbf{T}$ be an element in $M(m,k)$, then we can rewrite it as

$$\mathbf{T} := (T_1,\ldots,T_m),$$

where each T_i is the image of $i\in\{1,2,..,m\}$ under the map $\mathbf{T}$ and

$$T_i\in\Big\{(i_1,\ldots,i_k) : i_j\in[km] \text{ and } i_j\neq i_\ell,\ \forall\, 1\le j<\ell\le k\Big\}.$$

We also write $T_i = (T_{i,1},\ldots,T_{i,k})$ where $T_{i,j}$ is the j-th component of the k-tuple T_i. For example, when $m=3$ and $k=2$, then $\mathbf{T},\mathbf{T}',\mathbf{T}''$ defined as follows all belong to $M(3,2)$,

$$T_1=(1,2), T_2=(1,4), T_3=(2,4). \tag{35}$$

$$T_1'=(1,3), T_2'=(1,6), T_3'=(3,6). \tag{36}$$

$$T_1''=(1,2), T_2''=(1,4), T_3''=(5,6). \tag{37}$$

We say two maps $\mathbf{T},\widehat{\mathbf{T}}\in M(m,k)$ are *equivalent* if they differ by a permutation of $[km]$, i.e. by composing with a permutation they become the same map. We denote by $S(m,k)$ the set of all equivalence classes of $M(m,k)$. As an example, the $\mathbf{T}$ and $\mathbf{T}'$ defined in (35) and (36) are equivalent since the permutation (23)(46) brings $\mathbf{T}$ to $\mathbf{T}'$. But $\mathbf{T}''$ defined in (37) is neither equivalent to $\mathbf{T}$ nor $\mathbf{T}'$.

For any $\mathbf{T}\in M(m,k)$, we can construct a graph for it, which we call $\mathbf{T}$-*graph*. The $\mathbf{T}$-graph is constructed in two steps. Initially there are mk vertices in total, indexed by (i,j) for $1\le i\le m, 1\le j\le k$. First for each $1\le i\le m$ and $1\le j\le k-1$, we draw a black edge between $T_{i,j}$ and $T_{i,j+1}$. Then for any $(i,j)\neq(i',j')$ such that $T_{i,j}=T_{i',j'}$, we use a solid red edge to connect (i,j) and (i',j'). See Figure 1 for the graphical representations of $\mathbf{T}$, $\mathbf{T}'$ and $\mathbf{T}''$. One can see that if $\mathbf{T}$ is equivalent to $\widehat{\mathbf{T}}$, then $\mathbf{T}$-graph is the same as $\widehat{\mathbf{T}}$-graph, and vice versa. Consequently, each equivalence class in $S(m,k)$ can be identified with a $\mathbf{T}$-graph.

For $\mathbf{T}\in M(m,k)$, we define the size and the range of $\mathbf{T}$ as

$$|\mathbf{T}| := |\cup_{i=1}^m T_i|\,,\ \ \mathrm{Range}(\mathbf{T}) = \cup T_i.$$

For example, for the $\mathbf{T}$ defined in (35) we have $|\mathbf{T}|=3$ and $\mathrm{Range}(\mathbf{T})=\{1,2,4\}$.

For notational simplicity, for a collection of indices $\mathbf{t}=(t_1,\ldots,t_k)$, we define $f(\mathbf{t}) := f(x_{t_1},x_{t_2},\ldots,x_{t_k})$; by an abuse of nation, for $\mathbf{T}=(T_1,\ldots,T_m)$, we set

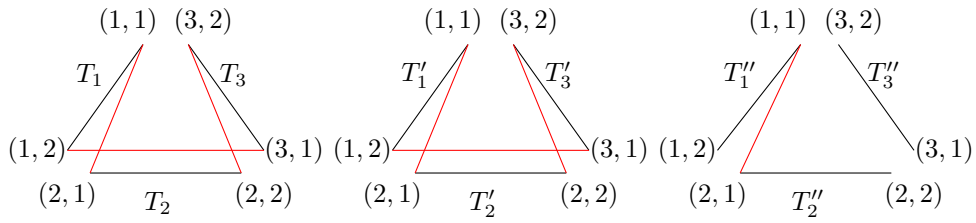

FIGURE 1. Graphical view of $\mathbf{T}$ (left), $\mathbf{T}'$ (middle) and $\mathbf{T}''$ (right)

$f(\mathbf{T}) = \Pi_{i=1}^{m} f(T_i)$; and we write $d\mathbf{x}$ as the volume element involved in the integration. By the definition of the determinantal point process, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{(x_1,\ldots,x_k)\in\Phi_*^k} f(x_1,\ldots,x_k)\right)^m\right] \\
&= \sum_{\mathbf{T}\in S(m,k)} \int_{\mathcal{X}^{|\mathbf{T}|}} f(\mathbf{T}) \det\Big(K(x_i,x_j)_{i,j\in \mathrm{Range}(\mathbf{T})}\Big) d\mathbf{x} \qquad (38)\\
&= \sum_{\mathbf{T}\in S(m,k)} \sum_{\sigma\in \mathrm{Sym}(\mathrm{Range}(\mathbf{T}))} \int_{\mathcal{X}^{|\mathbf{T}|}} f(\mathbf{T})\mathrm{sgn}(\sigma)\Pi_{q\in\mathrm{Range}(\mathbf{T})} K(x_q, x_{\sigma(q)}) d\mathbf{x}.
\end{aligned}$$

Here, for any set A, $\mathrm{Sym}(A)$ is the set of all permutations of the elements in A, and $\mathrm{sgn}(\sigma)$ is the sign of the permutation σ.

For any $\mathbf{T}$ and $\sigma \in \mathrm{Sym}(\mathrm{Range}(\mathbf{T}))$ we can further construct a $(\mathbf{T},\sigma)$-*graph* G by adding dotted red edges to the $\mathbf{T}$-graph. Specifically, for any $T_{i,j} \neq T_{i',j'}$, we add a dotted red edge between two vertices (i,j) and (i',j') if $\sigma(T_{i,j}) = T_{i',j'}$ or $\sigma(T_{i',j'}) = T_{i,j}$. We say the pair $(\mathbf{T},\sigma)$ is *connected* if the $(\mathbf{T},\sigma)$-graph is connected.

For example, for $\mathbf{T}$ defined in (35), the $(\mathbf{T},\sigma)$-graph G is connected for any $\sigma \in \mathrm{Sym}(\{1,2,4\})$ because the $\mathbf{T}$-graph itself is already connected. On the other hand, for $\mathbf{T}''$ defined in (37), if $\sigma = id$ (the identity in the permutation group), then the $(\mathbf{T}'',\sigma)$-graph has two components. However, for $\sigma = (15) \in \mathrm{Sym}(\{1,2,4,5,6\})$ the $(\mathbf{T}'',\sigma)$-graph becomes connected.

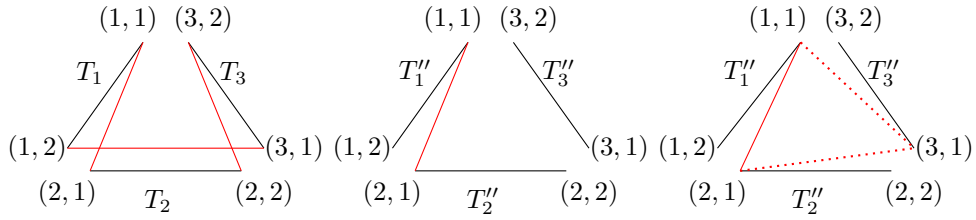

FIGURE 2. $\mathbf{T}$ in (35), $\sigma = id$ (left); $\mathbf{T}''$ in (37), $\sigma = id$ (middle); $\mathbf{T}''$ in (37), $\sigma = (15)$ (right).

If a $(\mathbf{T},\sigma)$-graph G has ℓ connected components, then G naturally induces a partition R of $[m]$ into ℓ disjoint sets $\{R_1,\ldots,R_\ell\}$. For $1 \le j \le \ell$, we set

$$H_j := \cup_{i\in R_j} T_i,\ \ \sigma_j = \text{the restriction of } \sigma \text{ to } H_j.$$

Let $f(\mathbf{T}|_{R_j}) = \prod_{i\in R_j} f(T_i)$. Then for the integral

$$\int_{\mathcal{X}^{|\mathbf{T}|}} f(T_1)f(T_2)\dots f(T_m)\mathrm{sgn}(\sigma)\Pi_{q\in\mathrm{Range}(\mathbf{T})}K(x_q,x_{\sigma(q)})d\mathbf{x},$$

we can split it into a product of exactly ℓ integrals

$$\prod_{j=1}^{\ell}\left(\int_{\mathcal{X}^{|H_j|}}\mathrm{sgn}(\sigma_j)f(\mathbf{T}|_{R_j})\prod_{q\in H_j}K_n(x_q,x_{\sigma_j(q)})dx_q\right).$$

For any integer-valued r, we define

$$\begin{aligned}\mathcal{C}(r) :=&\Big\{(\mathbf{T},\sigma) : \mathbf{T}\in S(r,k), \sigma\in \mathrm{Sym}(\mathrm{Range}(\mathbf{T})),\\ &(\mathbf{T},\sigma)\text{-graph is connected}\Big\}.\end{aligned} \tag{39}$$

The definition of $\{R_1,\dots,R_\ell\}$ implies that, for each $1\le j\le \ell$, the pair $(\mathbf{T}|_{R_j},\sigma_j)$ is in $\mathcal{C}(|R_j|)$. Therefore, we have

$$\begin{aligned}&\sum_{\mathbf{T}\in S(m,k)}\sum_{\sigma\in\mathrm{Sym}(\mathrm{Range}(\mathbf{T}))}\int_{\mathcal{X}^{|\mathbf{T}|}}f(\mathbf{T})\mathrm{sgn}(\sigma)\prod_{q\in\mathrm{Range}(\mathbf{T})}K(x_q,x_{\sigma(q)})d\mathbf{x}\\ =&\sum_{R=\{R_1,\dots,R_\ell\}\in\Pi(m)}\prod_{j=1}^{\ell}\left(\sum_{(\mathbf{T},\sigma)\in\mathcal{C}(|R_j|)}\mathfrak{Int}(f,(\mathbf{T},\sigma))\right),\end{aligned} \tag{40}$$

where

$$\mathfrak{Int}(f,(\mathbf{T},\sigma)) := \int_{\mathcal{X}^{|\mathbf{T}|}}\left(f(\mathbf{T})\mathrm{sgn}(\sigma)\prod_{q\in\mathrm{Range}(\mathbf{T})}K(x_q,x_{\sigma(q)})\right)d\mathbf{x}.$$

Combining (31) (38) and (40), we obtain the following formula for the cumulants of multivariate linear statistics of general determinantal point processes.

Lemma 1.

$$\begin{aligned}&Q_m\left(\sum_{(x_1,\dots,x_k)\in\Phi_*^k}f(x_1,\dots,x_k)\right)\\ =&\sum_{(\mathbf{T},\sigma)\in\mathcal{C}(m)}\int_{\mathcal{X}^{|\mathbf{T}|}}f(\mathbf{T})\mathrm{sgn}(\sigma)\prod_{q\in\mathrm{Range}(\mathbf{T})}K(x_q,x_{\sigma(q)})d\mathbf{x}.\end{aligned} \tag{41}$$

For $k=1$, (41) gives the following well-known formula (Formula (2.7) in [8]),

$$\begin{aligned}&Q_m\left(\sum_{x\in\Phi}f(x)\right)\\ =&\sum_{\ell=1}^{m}\sum_{(n_1,\dots,n_\ell):\sum_{j=1}^{\ell}n_j=m,n_j\ge1,\forall j}\frac{(-1)^{\ell-1}}{\ell}\frac{m!}{n_1!\dots n_\ell!}\\ &\int_{\mathcal{X}^\ell}f^{n_1}(x_1)\cdots f^{n_\ell}(x_\ell)K(x_1,x_2)\cdots K(x_{\ell-1},x_\ell)K(x_\ell,x_1)d\mathbf{x}.\end{aligned} \tag{42}$$

Indeed, by the definition of $S(m,1)$, each $\mathbf{T}\in S(m,1)$ corresponds to one way of assigning m different balls into ℓ indistinguishable urns for some ℓ. Hence the $\mathbf{T}$-graph itself has ℓ components and also partitions the set $[m]$ into ℓ components.

Thus, to ensure the $(\mathbf{T}, \sigma)$-graph is connected, different components have to be linked through $\sigma \in \mathrm{Sym}(\mathrm{Range}(\mathbf{T}))$, which implies that σ has to be a cyclic permutation of length ℓ. As an example, suppose $k = 1$, $m = 5$ and $\mathbf{T} = \{1, 2, 3, 3, 3\}$, then $|\mathbf{T}| = 3$ and σ has to be (123) or (132) to obtain a connected $(\mathbf{T}, \sigma)$-graph.

For later reference, we introduce a few more concepts.

Definition 1. *For a $(\mathbf{T}, \sigma)$-graph, we say $T_{i,j}$ is a* connection point *if at least one of the two conditions are satisfied:*

- $\sigma(T_{i,j}) \notin T_i$.
- *There exists an $i' \neq i$ such that $T_{i,j} \in T_{i'}$.*

Equivalently, using the graphical representation of a $(\mathbf{T}, \sigma)$-graph, $T_{i,j}$ is a connection point if (i, j) is connected to some vertex in $\{(i', j') : i' \neq i, 1 \leq j' \leq k\}$ by a red edge, either solid or dotted.

Note that, if the $(\mathbf{T}, \sigma)$-graph is connected, then for each i, there must exist at least one connection point $T_{i,j}$.

Definition 2. *We say a $(\mathbf{T}, \sigma)$ pair is* reducible *if its $(\mathbf{T}, \sigma)$-graph is connected and there exists an $i \in [m]$ and a $j \in [k]$ such that*

- *$T_{i,j}$ is the only connection point in T_i.*
- *$\sigma(x) = x, \forall x \in T_i - \{T_{i,j}\}$.*

If the above two conditions hold, then we say the $(\mathbf{T}, \sigma)$-graph breaks *at $T_{i,j}$ and $T_{i,j}$ is a* break point*. Equivalently, $(\mathbf{T}, \sigma)$-graph is reducible if it is connected and there exists some (i, j) which is the only vertex in $\{(i, j) : 1 \leq j \leq k\}$ that can have red edge(s) connecting with other vertices. We say a $(\mathbf{T}, \sigma)$ pair is* irreducible *if it is not reducible.*

We define $\mathfrak{I}(m)$ to be the set of all $(\mathbf{T}, \sigma) \in \mathcal{C}(m)$ that are irreducible, i.e.,

$$\mathfrak{I}(m) := \{(\mathbf{T}, \sigma) \in \mathcal{C}(m) : (\mathbf{T}, \sigma) \text{ is irreducible}\}. \tag{43}$$

An example of the reducible graph is given by the right panel of Figure 2, while the left and the middle ones in Figure 2 are irreducible.

Definition 3. *We say a $(\mathbf{T}, \sigma) \in \mathfrak{I}(m)$ is* circle-like *if for each $1 \leq i \leq m$, there are exactly two distinct numbers $1 \leq i_1 \neq i_2 \leq k$ such that each of (i, i_1) and (i, i_2) has exactly one red edge and the red edge is connected to a vertex in $\{(i', j') : i' \neq i, 1 \leq j' \leq k\}$, and all other vertices, i.e., those not in the set $\{(i, i_1), (i, i_2) : 1 \leq i \leq m\}$, have no red edge.*

The following proposition explains the name 'circle-like'.

Proposition 1. *Let $(\mathbf{T}, \sigma)$ be circle-like. Then there exists a cyclic permutation p of $\{1, \ldots, m\}$ such that, for each $1 \leq i \leq m$, there exist two distinct indices i_1 and i_2 and that (i, i_2) is connected with $(p(i), p(i)_1)$ with a red edge.*

Proof. Note that, by the definition of being circle-like, if we contract all vertices in $\{(i, j) : 1 \leq j \leq k\}$ into a single vertex (and give it label i), then we will obtain a connected graph with m vertices such that each vertex has degree 2, which is then necessarily a circle of size m. Fix a direction of the circle, suppose the label of these vertices are $a_1, \ldots, a_m$. Then we can define a permutation p such that $p(a_i) = a_{i+1}$ where $a_{m+1} := a_1$. In addition, by reordering i_1 and i_2 for each $1 \leq i \leq m$ if needed, we can assume that $(a_i, (a_i)_2)$ is connected with $(a_{i+1}, (a_{i+1})_1)$ for all $1 \leq i \leq m$ with a red edge. This completes the proof. □

As a remark, we will see that in the proof of Theorem 3 for the degenerate case, the collections of the cycle-like $(\mathbf{T}, \sigma)$-graph will provide the leading order term for the cumulants of multivariate linear statistics, which will eventually yield the 2nd order Wiener chaos.

3. Properties of the spectral projection kernel

In this section, we first review some basic facts for the spectral projection kernel. Then we will derive several integral lemmas which provide the key estimates to prove the main results.

3.1. **Preliminaries.** It's well-known that the kernel for the spectral orthogonal projection $K_n : L^2(S^d) \to \mathcal{H}_n(S^d)$ satisfies (Theorem 2.9 in [2])

$$K_n(x,y) = \frac{k_n}{s_d} P_n(\cos \mathrm{d}(x,y)) = \frac{k_n}{s_d} P_n(x \cdot y), \tag{44}$$

where $\mathrm{d}(x,y) \in [0,\pi]$ is the geodesic distance which is the angle between the vectors $x, y \in S^d$, P_n is the Legendre polynomial of degree n in d dimension, k_n is the dimension of $\mathcal{H}_n$ given in (4) and $s_d = 2\pi^{\frac{d+1}{2}}/\Gamma(\frac{d+1}{2})$ is the surface area of S^d. Since both x and y are on the unit sphere, then we can rewrite $\cos \mathrm{d}(x,y) = x \cdot y$ as the inner product between x and y.

We also write

$$P_n(x,y) := P_n(\cos \mathrm{d}(x,y)) = P_n(x \cdot y).$$

By the fact that $P_n(1) = 1$ [2], one has the identity

$$K_n(x,x) = \frac{k_n}{s_d}. \tag{45}$$

The kernel $K_n(x,y)$ satisfies the reproducing property,

$$\int_{S^d} K_n(x_1,x_2) K_n(x_2,x_3) dx_2 = K_n(x_1,x_3). \tag{46}$$

When $x_1 = x_3$, (46) reads,

$$\int_{S^d} K_n^2(x_1,x_2) dx_2 = \frac{k_n}{s_d}, \tag{47}$$

and thus we have

$$\int_{(S^d)^2} K_n^2(x_1,x_2) dx_1 dx_2 = k_n. \tag{48}$$

For P_n, two basic properties are [2],

$$P_n(x) = (-1)^n P_n(-x) \tag{49}$$

and

$$|P_n(x)| \le 1, \ \forall x \in [-1,1]. \tag{50}$$

By (45) and the reproducing property (46), we obtain that

$$\int_{S^d} P_n(x_1,x_2) P_n(x_2,x_3) dx_2 = \left(\frac{k_n}{s_d}\right)^{-1} P_n(x_1,x_3), \tag{51}$$

and by (47), we have

$$\int_{S^d} P_n^2(x_1,x_2) dx_2 = \left(\frac{k_n}{s_d}\right)^{-1}. \tag{52}$$

For $0 \le \theta \le \pi/2$, one has the Hilb's asymptotics for the Legendre polynomials (by taking $\alpha = \beta = \frac{d-2}{2}$ in [10, Theorem 8.21.12]),

$$\begin{aligned} P_n(\cos\theta) =& \Gamma\left(\frac{d}{2}\right)\left(\frac{\theta}{\sin\theta}\right)^{\frac{1}{2}}\left(\frac{1}{2}(n+\frac{d-1}{2})\sin\theta\right)^{-\frac{d-2}{2}} J_{\frac{d-2}{2}}\left((n+\frac{d-1}{2})\theta\right) \\ &+ R_n(\theta), \end{aligned} \tag{53}$$

where $J_{\frac{d-2}{2}}$ is the Bessel function of order $\frac{d-2}{2}$. And the error term satisfies the estimates:

$$R_n(\theta) = \begin{cases} \theta^2 O(1) & 0 \le \theta \le cn^{-1} \ , \\ \theta^{\frac{3-d}{2}} O(n^{-\frac{1+d}{2}}) & cn^{-1} \le \theta \le \pi/2 \ , \end{cases}$$

where c is some constant independent of n.

For the Bessel function, $J_{\frac{d-2}{2}}$ is bounded on the positive real line and has the expansion (Formula (1.71.1) in [10]),

$$J_{\frac{d-2}{2}}(x) = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!\Gamma(j+\frac{d}{2})}\left(\frac{x}{2}\right)^{2j+\frac{d-2}{2}}. \tag{54}$$

Furthermore, it admits the asymptotic expansion (Formula (1.71.7) in [10]),

$$J_{\frac{d-2}{2}}(x) = \sqrt{\frac{2}{\pi x}}\cos\left(x-(d-1)\frac{\pi}{4}\right) + o(x^{-1}) \quad \text{as } x \to +\infty. \tag{55}$$

Now we define a function $p_n(\theta)$ for $\theta \in [0,\pi]$ as follows. For $0 \le \theta \le \pi/2$, we define

$$\begin{aligned} p_n(\theta) :=& \Gamma\Big(\frac{d}{2}\Big)\left(\frac{\theta}{\sin\theta}\right)^{1/2}\left(\frac{1}{2}(n+\frac{d-1}{2})\sin\theta\right)^{-(d-2)/2} \\ &\times \sqrt{\frac{2}{\pi(n+(d-1)/2)\theta}}\cos\Big((n+(d-1)/2)\theta-(d-1)\frac{\pi}{4}\Big) \\ =& \Gamma\Big(\frac{d}{2}\Big)\Big(\frac{2^{d-1}}{\pi}\Big)^{1/2}\Big(n+(d-1)/2\Big)^{-(d-1)/2}\big(\sin\theta\big)^{-(d-1)/2} \\ &\times \cos\Big((n+(d-1)/2)\theta-(d-1)\frac{\pi}{4}\Big); \end{aligned} \tag{56}$$

for $\pi/2 < \theta \le \pi$, we define

$$p_n(\theta) := (-1)^n p_n(\pi-\theta).$$

Combining (53) and (55), for $0 \le \theta \le \pi$, we have the estimates,

$$|P_n(\cos\theta) - p_n(\theta)| \le C\left(\min\{n\theta, n(\pi-\theta)\}^{-d/2} \wedge 1\right), \tag{57}$$

$$|P_n(\cos\theta)| \le C\left(\min\{n\theta, n(\pi-\theta)\}^{-(d-1)/2} \wedge 1\right), \tag{58}$$

and

$$|p_n(\theta)| \le C\left(\min\{n\theta, n(\pi-\theta)\}^{-(d-1)/2} \wedge 1\right). \tag{59}$$

3.2. **Integral estimates.** Now we will prove several lemmas involving the integrals of the kernel K_n. They will be one of the main technical ingredients in the proofs of our main results.

We will use the spherical coordinate system $(\theta, \phi_1, \ldots, \phi_{d-1})$ for S^d, where $\theta, \phi_1, \ldots, \phi_{d-2}$ range over [0, π] and ϕ_{d-1} ranges over [0,2π]. Here, θ is the arc length from the point (θ, ϕ) to the origin of the coordinate system. For simplicity, we will use ϕ as a shorthand for $(\phi^1, \ldots, \phi^{d-1})$, and thus the range of ϕ is $\Omega := [0, \pi]^{d-2} \times [0, 2\pi]$. Then the volume element for S^d with respect to the standard round metric is

$$dx = \widehat{J}(\theta, \phi) d\theta d\phi,$$

where

$$\widehat{J}(\theta, \phi) = \sin^{d-1}(\theta) \sin^{d-2}(\phi_1) \cdots \sin(\phi_{d-2}).$$

We define

$$J(\phi) := \sin^{d-2}(\phi_1) \cdots \sin(\phi_{d-2}),$$

and thus we can rewrite

$$dx = \sin^{d-1}(\theta) J(\phi) d\theta d\phi.$$

The first lemma concerns the integration of a function against K_n^2.

Lemma 2. *For any bounded function $f(x, y)$, we have*

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{k_n} \int_{S^d} \int_{S^d} f(x,y) K_n^2(x,y) dxdy \\ =&\frac{2^{d-1}}{\Gamma(d)\pi} \Big(\frac{\Gamma\big(\frac{d}{2}\big)}{s_d} \Big)^2 \int_{S^d} \int_0^\pi \int_\Omega f(x, x + (\theta, \phi)) J(\phi) d\phi d\theta dx. \qquad (60) \\ =&\frac{2^{d-1}}{\Gamma(d)\pi} \Big(\frac{\Gamma\big(\frac{d}{2}\big)}{s_d} \Big)^2 \int_{S^d} \int_{S^d} \frac{f(x,y)}{\sin^{d-1}(\arccos(x \cdot y))} dxdy. \end{aligned}$$

The next two lemmas give upper bounds on the integration of the product of several $K_n's$.

Lemma 3. *For any $r \in \mathbb{N}, r \geq 2$,*

$$\int_{(S^d)^r} \Big| \prod_{i=1}^r K_n(x_i, x_{i+1}) \Big| dx_1 \cdots dx_r = O(n^{\frac{(d-1)r}{2}}), \qquad (61)$$

where x_{r+1} is set to be x_1. Equivalently,

$$\int_{(S^d)^r} \Big| \prod_{i=1}^r P_n(x_i, x_{i+1}) \Big| dx_1 \cdots dx_r = O(n^{-\frac{(d-1)r}{2}}). \qquad (62)$$

Lemma 4. *For any $r \in \mathbb{N}, r \geq 3$ and bounded measurable function h of r variables,*

$$\int_{(S^d)^r} h(x_1, \ldots, x_r) \prod_{i=1}^r P_n(x_i, x_{i+1}) dx_1 \cdots dx_r = o(n^{-\frac{(d-1)r}{2}}), \qquad (63)$$

where x_{r+1} is set to be x_1.

We now give the proofs of Lemmas 2-4.

Proof of Lemma 2. By the boundedness of f, without loss of generality, we may assume that f is nonnegative. For $x, y \in S^d$, we build a spherical coordinate system (θ, ϕ) with x being the north pole and write y as $x + (\theta, \phi)$. By the facts that $K_n(x, y) = k_n P_n(\cos\theta)/s_d$ and $|P_n \cos(\theta)| = |P_n \cos(\pi - \theta)|$, we have

$$\begin{aligned}
&\int_{S^d}\int_{S^d} f(x,y)K_n^2(x,y)dxdy \\
=&\Big(\frac{k_n}{s_d}\Big)^2 \int_{S^d}\int_0^{\pi}\int_{\Omega} f(x, x+(\theta,\phi))P_n(\cos\theta)^2\widehat{J}(\theta,\phi)d\phi d\theta dx \\
=&\Big(\frac{k_n}{s_d}\Big)^2 \left(\int_{S^d}\int_0^{\frac{\pi}{2}}\int_{\Omega} f(x, x+(\theta,\phi))P_n(\cos\theta)^2\widehat{J}(\theta,\phi)d\phi d\theta dx \right. \\
&\qquad \left. + \int_{S^d}\int_0^{\frac{\pi}{2}}\int_{\Omega} f(x, x+(\pi-\theta,\phi))P_n(\cos\theta)^2\widehat{J}(\theta,\phi)d\phi d\theta dx\right) \\
:=&\Big(\frac{k_n}{s_d}\Big)^2 (I_1 + I_2).
\end{aligned} \tag{64}$$

We will analyze I_1 and I_2 by a series of approximations. We only give details for I_1, and I_2 follows from the same arguments. By Hilb's asymptotic (53), one has

$$\begin{aligned}
P_n(\cos\theta)^2 =&\Gamma\Big(\frac{d}{2}\Big)^2 \left(\frac{1}{2}(n+\frac{d-1}{2})\sin\theta\right)^{-(d-2)} \Big(\frac{\theta}{\sin\theta}\Big) J_{\frac{d-2}{2}}\left((n+\frac{d-1}{2})\theta\right)^2 \\
&+ \widehat{R}_n(\theta),
\end{aligned} \tag{65}$$

where

$$\widehat{R}_n(\theta) = \begin{cases} \theta^2 O(1) & 0 \le \theta \le c/n\ , \\ \theta^{2-d} O(n^{-d}) & c/n \le \theta \le \pi/2\ . \end{cases}$$

We now define

$$I_3 = \int_{S^d}\int_0^{\frac{\pi}{2}}\int_{\Omega} \theta J_{\frac{d-2}{2}}\Big((n+\frac{d-1}{2})\theta\Big)^2 f(x, x+(\theta,\phi))J(\phi)d\phi d\theta dx. \tag{66}$$

By (64) and (65) there exists $C > 0$ such that

$$\left| I_1 - \Gamma\Big(\frac{d}{2}\Big)^2 \left(\frac{1}{2}(n+\frac{d-1}{2})\right)^{-(d-2)} I_3 \right| \le Cn^{-d}. \tag{67}$$

By (55), for any $\epsilon \in (0,1)$, there exists an $M > 0$ large enough such that for $x > M$, we have

$$\begin{aligned}
&(1-\epsilon)\frac{2}{\pi x}\cos^2\left(x - (d-1)\frac{\pi}{4}\right) - x^{-\frac{3}{2}} \\
\le& J_{\frac{d-2}{2}}(x)^2 \\
\le&(1+\epsilon)\frac{2}{\pi x}\cos^2\left(x - (d-1)\frac{\pi}{4}\right) + x^{-\frac{3}{2}}.
\end{aligned} \tag{68}$$

Now we split I_3 into two terms,

$$\begin{aligned} I_3 =& \int_{S^d}\int_0^{M/n}\int_\Omega J_{\frac{d-2}{2}}\left((n+\frac{d-1}{2})\theta\right)^2 f(x,x+(\theta,\phi))\theta J(\phi)d\phi d\theta dx \\ &+\int_{S^d}\int_{M/n}^{\frac{\pi}{2}}\int_\Omega J_{\frac{d-2}{2}}\left((n+\frac{d-1}{2})\theta\right)^2 f(x,x+(\theta,\phi))\theta J(\phi)d\phi d\theta dx \\ :=& I_4+I_5. \end{aligned} \tag{69}$$

For I_4, by the boundedness of f and $J_{\frac{d-2}{2}}$, there exists some $C>0$ such that

$$|I_4|\le CM^2/n^2. \tag{70}$$

For I_5, it holds trivially that $(n+\frac{d-1}{2})\theta > M$ for $\theta > M/n$. Hence, we can apply the estimates (68) for $J_{\frac{d-2}{2}}((n+\frac{d-1}{2})\theta)$. We set

$$\begin{aligned} I_6 =& \int_{S^d}\int_{M/n}^{\frac{\pi}{2}}\int_\Omega \frac{2}{\pi(n+(d-1)/2)\theta} f(x,x+(\theta,\phi)) \\ &\times\cos^2\Big((n+\frac{d-1}{2})\theta-(d-1)\frac{\pi}{4}\Big)\theta J(\phi)d\phi d\theta dx. \end{aligned} \tag{71}$$

Combining (68), (71) and the following estimate

$$\int_{S^d}\int_{M/n}^{\frac{\pi}{2}}\int_\Omega\left((n+\frac{d-1}{2}\theta)\right)^{-3/2} f(x,x+(\theta,\phi))\theta J(\phi)d\phi d\theta dx \le Cn^{-3/2},$$

we have that

$$(1-\epsilon)I_6 - Cn^{-\frac{3}{2}} \le I_5 \le (1+\epsilon)I_6 + Cn^{-\frac{3}{2}}. \tag{72}$$

By Riemann-Lebesgue lemma, for any fixed x and ϕ, one has

$$\begin{aligned} &\lim_{n\to\infty}\int_{M/n}^{\frac{\pi}{2}} f(x,x+(\theta,\phi))\cos^2\Big((n+\frac{d-1}{2})\theta-(d-1)\frac{\pi}{4}\Big)d\theta \\ =&\lim_{n\to\infty}\int_{M/n}^{\frac{\pi}{2}} f(x,x+(\theta,\phi))\Big(\frac{1}{2}+\frac{\cos((2n+d-1)\theta-(d-1)\frac{\pi}{2})}{2}\Big)d\theta \\ =&\frac{1}{2}\int_0^{\frac{\pi}{2}} f(x,x+(\theta,\phi))d\theta. \end{aligned} \tag{73}$$

Therefore, the bounded convergence theorem implies that

$$\begin{aligned} &\lim_{n\to\infty}\int_{S^d}\int_\Omega\int_{M/n}^{\frac{\pi}{2}} f(x,x+(\theta,\phi))\cos^2\Big((n+\frac{d-1}{2})\theta-\frac{\pi}{4}\Big)J(\phi)d\theta d\phi dx \\ =&\frac{1}{2}\int_{S^d}\int_\Omega\int_0^{\frac{\pi}{2}} f(x,x+(\theta,\phi))J(\phi)d\theta d\phi dx. \end{aligned} \tag{74}$$

This implies that

$$\lim_{n\to\infty} nI_6 = \frac{1}{\pi}\int_{S^d}\int_\Omega\int_0^{\frac{\pi}{2}} f(x,x+(\theta,\phi))J(\phi)d\theta d\phi dx := I_7. \tag{75}$$

Now, combining (69), (70) and (72) and (75), we have

$$(1-\epsilon)I_7 \le \liminf_{n\to\infty} nI_3 \le \limsup_{n\to\infty} nI_3 \le (1+\epsilon)I_7. \tag{76}$$

Since (76) holds for all $\epsilon \in (0,1)$ while I_3 and I_7 don't depend on ϵ, by sending $\epsilon \to 0$, we have

$$\lim_{n\to\infty} nI_3 = I_7. \tag{77}$$

Combining (67) and (77), we have

$$\lim_{n\to\infty} n^{d-1} I_1 = \Gamma\Big(\frac{d}{2}\Big)^2 2^{d-2} I_7.$$

By the same argument with θ replaced by $\pi - \theta$, we get a similar limit

$$\lim_{n\to\infty} n^{d-1} I_2 = \Gamma\Big(\frac{d}{2}\Big)^2 2^{d-2} I_8,$$

where I_8 is defined similarly to I_7 as

$$I_8 = \frac{1}{\pi}\int_{S^d}\int_{\Omega}\int_{\frac{\pi}{2}}^{\pi} f(x, x+(\theta,\phi))J(\phi)d\theta d\phi dx.$$

By the fact $k_n \sim 2n^{d-1}/\Gamma(d)$, we get

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{k_n}\int_{S^d}\int_{S^d} f(x,y)K_n^2(x,y)dxdy \\
=&\lim_{n\to\infty}\left(\frac{k_n}{s_d^2}\right)(I_1+I_2) \\
=&\lim_{n\to\infty}\frac{2n^{d-1}}{\Gamma(d)}\frac{1}{s_d^2}\Gamma\Big(\frac{d}{2}\Big)^2 2^{d-2} n^{-(d-1)}(I_7+I_8) \\
=&\frac{\Gamma\Big(\frac{d}{2}\Big)^2 2^{d-1}}{\pi\Gamma(d)s_d^2}\int_{S^d}\int_0^{\pi}\int_{\Omega} f(x,x+(\theta,\phi))J(\phi)d\phi d\theta dx.
\end{aligned} \tag{78}$$

This completes the proof of Lemma 2.

□

As a remark, the proof of Lemma 2 actually shows that for almost all x, we have

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{k_n}\int_{S^d} f(x,y)K_n^2(x,y)dy \\
=&\frac{2^{d-1}}{\Gamma(d)\pi}\Big(\frac{\Gamma\big(\frac{d}{2}\big)}{s_d}\Big)^2\int_{S^d}\frac{f(x,y)}{\sin^{d-1}(\arccos(x\cdot y))}dy.
\end{aligned} \tag{79}$$

Proof of Lemma 3. To prove Lemma 3, we recall (58) where we have

$$P_n(x,y) \le Cn^{-\frac{d-1}{2}}(\min\{\mathrm{d}(x,y), \pi - \mathrm{d}(x,y)\})^{-\frac{d-1}{2}}. \tag{80}$$

For $x_1,\dots,x_d \in S^d$, let $\alpha_{i,j} = \mathrm{d}(x_i,x_j)$ be the geodesic distance which is the angle between x_i and x_j and let $\beta_{i,j} = \min\{\alpha_{i,j}, \pi - \alpha_{i,j}\}$. We now claim

$$\beta_{1,3} \le \beta_{1,2} + \beta_{2,3}. \tag{81}$$

To prove (81), we consider four possible cases.

- If $\alpha_{1,2} < \pi/2$ and $\alpha_{2,3} \le \pi/2$, then we have

$$\beta_{1,2} + \beta_{2,3} = \alpha_{1,2} + \alpha_{2,3} \ge \alpha_{1,3} \ge \beta_{1,3}.$$

  Here, the first inequality follows from triangle inequality.

- If $\alpha_{1,2} < \pi/2$ and $\alpha_{2,3} \geq \pi/2$, then by symmetry of the sphere, if we set $x_3' := -x_3$ (the reflection of x_3 through the origin of $\mathbb{R}^{d+1}$), we have

$$\beta_{1,2} + \beta_{2,3} = \alpha_{1,2} + \pi - \alpha_{2,3} = \mathrm{d}(x_1, x_2) + \mathrm{d}(x_2, x_3') \geq \mathrm{d}(x_1, x_3') \geq \beta_{1,3}.$$

- The case $\alpha_{1,2} \geq \pi/2$ and $\alpha_{2,3} < \pi/2$ can be analyzed similarly to the second case.
- If $\alpha_{1,2} \geq \pi/2$ and $\alpha_{2,3} \geq \pi/2$, then by setting $x_2' := -x_2$, we have

$$\beta_{1,2} + \beta_{2,3} = \mathrm{d}(x_1, x_2') + \mathrm{d}(x_2', x_3) \geq \mathrm{d}(x_1, x_3) \geq \beta_{1,3}.$$

The inequality (81) implies that

$$\beta_{1,2}\beta_{2,3} = \max\{\beta_{1,2}, \beta_{2,3}\} \min\{\beta_{1,2}, \beta_{2,3}\} \geq \frac{\beta_{1,3}}{2} \min\{\beta_{1,2}, \beta_{2,3}\},$$

which gives

$$\begin{aligned} (\beta_{1,2}\beta_{2,3})^{-(d-1)/2} &\leq C\beta_{1,3}^{-(d-1)/2} \min\{\beta_{1,2}, \beta_{2,3}\}^{-(d-1)/2} \\ &\leq C\beta_{1,3}^{-(d-1)/2} \left(\beta_{1,2}^{-(d-1)/2} + \beta_{2,3}^{-(d-1)/2}\right). \end{aligned} \tag{82}$$

By (80) and (82), for any fixed x_1 and x_3, we have

$$\begin{aligned} &\int_{S^d} \Big| P_n(x_1, x_2) P_n(x_2, x_3) \Big| dx_2 \\ \leq & C n^{-(d-1)} \int_{S^d} (\beta_{1,2}\beta_{2,3})^{-(d-1)/2} dx_2 \\ \leq & C n^{-(d-1)} \beta_{1,3}^{-(d-1)/2} \int_{S^d} \left(\beta_{1,2}^{-(d-1)/2} + \beta_{2,3}^{-(d-1)/2}\right) dx_2 \\ \leq & C n^{-(d-1)} \beta_{1,3}^{-(d-1)/2} \left(\int_0^\pi \beta_{1,2}^{-(d-1)/2} \sin^{d-1}(\alpha_{1,2}) d\alpha_{1,2} \right. \\ & \left. + \int_0^\pi \beta_{2,3}^{-(d-1)/2} \sin^{d-1}(\alpha_{2,3}) d\alpha_{2,3} \right) \\ \leq & C n^{-(d-1)} \beta_{1,3}^{-(d-1)/2}. \end{aligned} \tag{83}$$

Using (83) $r-2$ times to integrate out the variables $x_2, \ldots, x_{r-1}$, we get

$$\begin{aligned} &\int_{(S^d)^r} \Big| \prod_{i=1}^r P_n(x_i, x_{i+1}) \Big| dx_1 \cdots dx_r \\ \leq & C n^{-(d-1)r/2} \int_{S^d} \Big(\int_0^\pi \beta_{1,r}^{-(d-1)} \sin^{(d-1)}(\alpha_{1,r}) d\alpha_{1,r} \Big) dx_1 \\ \leq & C n^{-(d-1)r/2}. \end{aligned} \tag{84}$$

This proves Lemma 3.

□

Proof of Lemma 4. As in the proof of Lemma 3, let $\alpha_{i,i+1}$ be the angle between x_i and x_{i+1} and set $\beta_{i,i+1} = \min\{\alpha_{i,i+1}, \pi - \alpha_{i,i+1}\}$. Recall the function p_n defined in (56), by (57),

$$|P_n(x_i, x_{i+1}) - p_n(\alpha_{i,i+1})| = |P_n(\cos\alpha_{i,i+1}) - p_n(\alpha_{i,i+1})| \leq C(n\beta_{i,i+1})^{-d/2}. \tag{85}$$

We can write

$$h(x_1, \ldots, x_r) \Pi_{i=1}^r P_n(x_i, x_{i+1}) = h(x_1, \ldots, x_r) \Pi_{i=1}^r p_n(\alpha_{i,i+1}) + I_r \tag{86}$$

where the error term I_r is bounded from above as

$$\begin{aligned} I_r \leq & C\,|h(x_1,\dots,x_r)| \sum_{j=1}^{r} (n\beta_{j,j+1})^{-d/2} \Big(\prod_{i=1, i\neq j}^{r} (|P_n(\cos\alpha_{j,j+1})| + |p_n(\alpha_{j,j+1})|) \Big) \\ \leq & C n^{-\frac{(d-1)(r-1)}{2}} n^{-d/2} \sum_{j=1}^{r} \left(\beta_{j,j+1}^{-d/2} \left(\prod_{i=1, i\neq j}^{r} \beta_{i,i+1}^{-(d-1)/2} \right) \right). \end{aligned} \tag{87}$$

The first inequality is given by the estimate (85) together with the following elementary inequality: given $a_1,\dots,a_r,b_1,\dots,b_r \in \mathbb{R}$, one has

$$\left| \prod_{i=1}^{r} a_i - \prod_{i=1}^{r} b_i \right| \leqslant \sum_{j=1}^{r} |a_j - b_j| \left(\prod_{i=1, i\neq j}^{r} (|a_i| + |b_i|) \right).$$

The second inequality in (87) is given by the estimates (58) and (59).

By slightly modifying the proof of Lemma 3 we can show that

$$\int_{(S^d)^r} \beta_{j,j+1}^{-d/2} \left(\prod_{i=1, i\neq j}^{r} \beta_{i,i+1}^{-\frac{d-1}{2}} \right) dx_1 \cdots dx_r < \infty. \tag{88}$$

Combining (87) and (88), we get

$$\int_{(S^d)^r} |I_r|\, dx_1 \cdots dx_r \leq C n^{-\frac{(d-1)r}{2} - \frac{1}{2}} = o(n^{-\frac{(d-1)r}{2}}). \tag{89}$$

We define a function

$$g(x_1,\dots,x_r) := h(x_1,\dots,x_r) \prod_{i=1}^{r} \sin^{-(d-1)/2}(\beta_{i,i+1}). \tag{90}$$

The proof of Lemma 3 implies that the function $\Pi_{i=1}^{r} \sin^{-\frac{d-1}{2}}(\beta_{i,i+1})$ is integrable over $(S^d)^r$. On the other hand, by definition of p_n, we can write

$$\begin{aligned} & h(x_1,\dots,x_r) \prod_{i=1}^{r} p_n(\alpha_{i,i+1}) \\ = & (n + (d-1)/2)^{-(d-1)r/2} \times g(x_1,\dots,x_r) \\ & \times \prod_{i=1}^{r} \left((-1)^{n\mathbf{1}[\alpha_{i,i+1} > \pi/2]} \cos\left((n + \frac{d-1}{2})\beta_{i,i+1} - (d-1)\frac{\pi}{4} \right) \right). \end{aligned} \tag{91}$$

When computing the integration over $x_1,\dots,x_r$, we can build a spherical coordinate system (θ,ϕ) around x_2 and represent x_1 by $x_2 + (\theta,\phi)$. Here $\theta \in [0,\pi]$ and ϕ has $d-1$ components $\phi_1,\dots,\phi_{d-1}$. We claim that, for almost every (fixed) $\phi, x_2,\dots,x_r$, the integration of (91) over θ has the limit

$$\begin{aligned} \lim_{n\to\infty} \int_{S^d} & g(x_2 + (\theta,\phi), x_2, \dots, x_r) \times \\ & \prod_{i=1,r} \left((-1)^{n\mathbf{1}[\alpha_{i,i+1} > \pi/2]} \cos\left((n + \frac{d-1}{2})\beta_{i,i+1} - (d-1)\frac{\pi}{4} \right) \right) d\theta = 0. \end{aligned} \tag{92}$$

Assume (92) for the moment, by (91) and the dominated convergence theorem, we have

$$\int_{(S^d)^r} h(x_1,\dots,x_r)\prod_{i=1}^{r} p_n(\alpha_{i,i+1})dx_1\cdots dx_r = o(n^{-\frac{(d-1)r}{2}}). \tag{93}$$

Lemma 4 now follows from (86), (89) and (93). Hence it remains to prove (92).

To this end we first rewrite the product of the two $\cos(\cdots)$ factors in (92) as

$$\begin{aligned}&\frac{1}{2}\cos\left((n+\frac{d-1}{2})(\beta_{1,2}+\beta_{r,r+1})-(d-1)\frac{\pi}{2}\right)\\ &+\frac{1}{2}\cos\left((n+\frac{d-1}{2})(\beta_{1,2}-\beta_{r,r+1})\right).\end{aligned} \tag{94}$$

Under the spherical coordinate system, $\alpha_{1,2}=\theta$ so that $\beta_{1,2}=\min\{\theta,\pi-\theta\}$. Denote by (θ',ϕ') the coordinate of x_r in this system. By making an orthogonal transformation if necessary, we may assume that $\phi_1'=0$.

To compute $\beta_{r,r+1}$, note that

$$\cos\alpha_{r,r+1}=\cos\alpha_{r,1}=x_1\cdot x_r=\cos\theta\cos\theta'+\sin\theta\cos\phi_1\sin\theta'. \tag{95}$$

If neither θ' nor ϕ_1 is not equal to 0 or π, then $\alpha_{r,r+1}$, viewed as a function of θ, is continuously differentiable at all but finite many θ's, and satisfies

$$\left|\frac{d\alpha_{r,r+1}}{d\theta}\right|=\frac{|-\sin\theta\cos\theta'+\cos\theta\cos\phi_1\sin\theta'|}{\sqrt{1-(\cos\theta\cos\theta'+\sin\theta\cos\phi_1\sin\theta')^2}}<1.$$

Thus, $\beta_{1,2}\pm\beta_{r,r+1}$ is piecewise differentiable in θ with a nonzero derivative. The limit (92) now follows from (94) and (the proof of) the Riemann-Lebesgue lemma.

Note that (92) is not true for $r=2$ where the second $\cos(\cdots)$ factor in (94) is a constant, which further implies that the integration (92) may tend to some constant other than 0. Thus we need the assumption $r\geq 3$.

□

4. Proof of Theorem 1

In this section we prove Theorem 1 regarding the asymptotic expansion of the mean $\mathbb{E}(L_nf)$. By (1) and (3), we have

$$\mathbb{E}(L_nf)=\int_{(S^d)^k} f(x_1,\dots,x_k)\det\Big(K_n(x_i,x_j)_{1\leq i,j\leq k}\Big)dx_1\cdots dx_k \tag{96}$$

We can expand the determinant as

$$\begin{aligned}\det\Big(K_n(x_i,x_j)_{1\leq i,j\leq k}\Big)=&\prod_{i=1}^{k}K_n(x_i,x_i)-\sum_{1\leq i<j\leq k}K_n^2(x_i,x_j)\prod_{\ell\neq i,j}K_n(x_\ell,x_\ell)\\ &+\text{remainder term},\end{aligned}$$

where the remainder term (denoted by I_9) is the sum of $\mathrm{sgn}(\sigma)\Pi_{i=1}^{k}K_n(x_i,x_{\sigma(i)})$ over all $\sigma's\in\mathrm{Sym}(k)$ which are neither the identity nor a transposition (a permutation which exchanges two elements and keeps all others fixed). Using the cycle

decomposition of permutations, (44) and (50), we have the upper bound

$$\begin{aligned}|I_9| \leq & C\left(\frac{k_n}{s_d}\right)^k \Bigg(\sum_{\sigma=(i_1j_1)(i_2j_2)} P_n^2(x_{i_1},x_{j_1})P_n^2(x_{i_2},x_{j_2}) \\ & + \sum_{3\leq r\leq k} \sum_{\sigma=(i_1\cdots i_r)} \left|P_n(x_{i_1},x_{i_2})\cdots P_n(x_{i_{r-1}},x_{i_r})P_n(x_{i_r},x_{i_1})\right| \Bigg),\end{aligned} \tag{97}$$

where C is some constant depending on k.

Combining (52), the estimate $k_n = \Theta(n^{d-1})$, the boundedness of f and Lemma 3, we have the upper bound

$$\int_{(S^d)^k} |f(x_1,\ldots,x_k)|\,|I_9|\,dx_1\cdots dx_k \leq Cn^{(d-1)k}(n^{-2(d-1)} + n^{-3(d-1)/2}), \tag{98}$$

which gives the error term in (12). We also have

$$\begin{aligned}&\int_{(S^d)^k} f(x_1,\ldots,x_k) \\ &\times \Big(\prod_{i=1}^k K_n(x_i,x_i) - \sum_{1\leq i<j\leq k} K_n^2(x_i,x_j) \prod_{\ell\neq i,j} K_n(x_\ell,x_\ell)\Big)dx_1\cdots dx_k \\ =& \left(\frac{k_n}{s_d}\right)^k \int_{(S^d)^k} f(x_1,\ldots,x_k)dx_1\cdots dx_k \\ &- \left(\frac{k_n}{s_d}\right)^{k-2} \int_{(S^d)^2} \sum_{1\leq i<j\leq k} f_{i,j}(x,y)K_n^2(x,y)dxdy,\end{aligned} \tag{99}$$

where $f_{i,j}$ is the (i,j)-margin function of f as defined in (11). Applying Lemma 2 to (99), we will get the first two terms in (12), which finishes the proof of Theorem 1.

5. Proof of Theorem 2

5.1. **Univariate case.** The univariate linear statistics for determinantal point processes has been understood very well. The following result proved in [9] is particularly useful. Given a family of determinantal point processes with kernel K_n and measurable bounded univariate functions f_n with compact support (to ensure integrability), let $L_n f_n$ and $L_n|f_n|$ be the linear statistics of f_n and $|f_n|$, respectively. Suppose that

$$\mathrm{Var}(L_n f_n) \to \infty,\ \sup|f_n| = o(\mathrm{Var}(L_n f_n)^\epsilon),\ \mathbb{E}(L_n|f_n|) = O(\mathrm{Var}(L_n f_n)^\delta) \tag{100}$$

for any $\epsilon > 0$ and some $\delta > 0$, then one has the central limit theorem,

$$\frac{L_n f_n - \mathbb{E}(L_n f_n)}{\sqrt{\mathrm{Var}(L_n f_n)}} \xrightarrow{\mathrm{d}} N(0,1).$$

In our case, the integrability condition holds trivially as the test function is bounded and the underlying space S^d is compact. Thus, it remains to check the three conditions in (100) in order to to prove Theorem 2 for the univariate case.

Note that the variance of $L_n f$ is given by

$$\mathrm{Var}(L_n f) = \frac{1}{2}\int_{S^d}\int_{S^d}(f(x)-f(y))^2 K_n^2(x,y)dxdy. \tag{101}$$

By Lemma 2, one immediately has the limit,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{\mathrm{Var}(L_n f)}{k_n}\\
&=\frac{2^{d-2}}{\Gamma(d)\pi}\Big(\frac{\Gamma(\frac{d}{2})}{s_d}\Big)^2\int_{S^d}\int_0^\pi\int_\Omega (f(x)-f(x+(\theta,\phi)))^2 J(\phi)d\phi d\theta dx. \qquad (102)\\
&=\frac{2^{d-2}}{\Gamma(d)\pi}\Big(\frac{\Gamma(\frac{d}{2})}{s_d}\Big)^2\int_{S^d}\int_{S^d}\frac{(f(x)-f(y))^2}{\sin^{d-1}(\arccos(x\cdot y))}dxdy.
\end{aligned}$$

By definition (10), the 1-margin function is itself for $k=1$, i.e., $F(x)=f(x)$, and thus (102) gives the limit of variance in (14) for $k=1$. The assumption that $F(x)$ is not constant almost everywhere implies the first condition $\mathrm{Var}(L_n f_n)=\Theta(k_n)\to\infty$. The second condition is satisfied since f is bounded. The third condition is satisfied with $\delta=1$ by the fact that

$$\mathbb{E}(L_n|f|) = \int_{S^d}|f(x)|\,K_n(x,x)dx = \frac{k_n}{s_d}\int_{S^d}|f(x)|\,dx = \Theta(k_n).$$

This completes the proof of Theorem 2 for the univariate case.

5.2. **Multivariate case.** Now we prove Theorem 2 for the multivariate linear statistics. There are two steps in the proof. We will first derive the growth order of the variance $\mathrm{Var}(L_n f)=Q_2(L_n f)$, then we will prove $Q_m(L_n f)=o(Q_2(L_n f)^{\frac{m}{2}})$ for all fixed $m\geq 3$. This will imply the Gaussian limit for the multivariate linear statistics by the method of cumulants.

We first introduce a notation. Given the set A which is a collection of $(\mathbf{T},\sigma)$-graph, we define

$$Q_m(L_n f, A) := \sum_{(\mathbf{T},\sigma)\in A}\int_{(S^d)^{|\mathbf{T}|}} f(\mathbf{T})\mathrm{sgn}(\sigma)\Pi_{q\in\mathrm{Range}(\mathbf{T})}K(x_q,x_{\sigma(q)})d\mathbf{x}, \tag{103}$$

where $d\mathbf{x}$ is the volume element involved in the integration. With such notation, we have $Q_m(L_n f)=Q_m(L_n f,\mathcal{C}(m))$ by (41) (recall the definition of $\mathcal{C}(m)$ in (39)).

We first estimate $Q_2(L_n f)$, which is the variance $\mathrm{Var}(L_n f)$. We can split the expression for $Q_2(L_n f)$ into 3 parts:

$$Q_2(L_n f) = Q_m(L_n f,\mathcal{C}(2)) = Q_2(L_n f, A_1)+Q_2(L_n f, A_2)+Q_2(L_n f, A_3),$$

where A_1, A_2, A_3 are disjoint subsets of $\mathcal{C}(2)$ defined as follows:

$$\begin{aligned}
A_1 &= \{(\mathbf{T},\sigma)\in\mathcal{C}(2) : |\mathbf{T}|=2k, \sigma \text{ is a transposition, i.e. } \sigma=(ij) \text{ for some } i,j\},\\
A_2 &= \{(\mathbf{T},\sigma)\in\mathcal{C}(2) : |\mathbf{T}|=2k-1, \sigma=id\},\\
A_3 &= \mathcal{C}(2)-A_1-A_2.
\end{aligned}$$

Lemma 5. *We have the following two estimates.*

(1)

$$\begin{aligned} &Q_2(L_nf, A_1) + Q_2(L_nf, A_2) \\ &= \Big(\frac{k_n}{s_d}\Big)^{2k-2} \frac{k_n 2^{d-2}}{\Gamma(d)\pi} \Big(\frac{\Gamma(\frac{d}{2})}{s_d}\Big)^2 \int_{S^d}\int_{S^d} \frac{(F(x)-F(y))^2}{\sin^{d-1}(\arccos(x\cdot y))} dxdy \\ &\quad + o(n^{(d-1)(2k-1)}). \end{aligned} \tag{104}$$

(2) $Q_2(L_nf, A_3) = o\left(n^{(d-1)(2k-1)}\right)$.

The limit (14) now follows from Lemma 5. In particular, since F is not constant almost everywhere, we have the following estimate of the variance

$$Q_2(L_nf) = \Theta(n^{(d-1)(2k-1)}). \tag{105}$$

Proof of Lemma 5 . We first consider $Q_2(L_nf, A_1)$. If $|\mathbf{T}| = 2k$, then $\mathbf{T}$ has to be $((1,\dots,k),(k+1,\dots,2k))$. Pick any $1 \le i \le k$ and $k+1 \le j \le 2k$. Then for such $\mathbf{T}$ and σ we have

$$\begin{aligned} Q_2(L_nf, (\mathbf{T},\sigma)) &= -\int_{(S^d)^{2k}} f(x_1,\dots,x_k) f(x_{k+1},\dots,x_{2k}) \Big(\frac{k_n}{s_d}\Big)^{2k-2} K_n^2(x_i,x_j) d\mathbf{x} \\ &= -\Big(\frac{k_n}{s_d}\Big)^{2k-2} \int_{(S^d)^2} f_i(x_i) f_{j-k}(x_j) K_n^2(x_i,x_j) dx_i dx_j, \end{aligned}$$

where the second equality is given by the definition of the i-margin function f_i in (10). Summing over all i, j, we see that $Q_2(L_nf, A_1)$ is equal to

$$\begin{aligned} &-\Big(\frac{k_n}{s_d}\Big)^{2k-2} \int_{(S^d)^2} \left(\sum_{i=1}^k f_i(x)\right)\left(\sum_{i=1}^k f_i(y)\right) K_n^2(x,y) dxdy \\ =&-\left(\frac{k_n}{s_d}\right)^{2k-2} \int_{(S^d)^2} F(x)F(y) K_n^2(x,y) dxdy. \end{aligned} \tag{106}$$

Now we consider A_2. Since $\mathbf{T} \in S(2,k)$ and $|\mathbf{T}| = 2k-1$, $\mathbf{T}$ has to satisfy $|T_1 \cap T_2| = 1$. The number of ways to choose 1 location in T_1 and 1 location in T_2 are both k. Therefore, $Q_2(L_nf, A_2)$ equals

$$\begin{aligned} &\sum_{i=1}^k \sum_{j=k+1}^{2k} \Big(\frac{k_n}{s_d}\Big)^{2k-1} \int_{(S^d)^{2k-1}} f(x_1,\dots,x_k) f(x_{k+1},\dots,x_{j-1},x_i,x_j,\dots,x_{2k-1}) d\mathbf{x} \\ =&\sum_{i=1}^k \sum_{j=k+1}^{2k} \Big(\frac{k_n}{s_d}\Big)^{2k-1} \int_{S^d} f_i(x) f_{j-k}(x) dx \\ =&\Big(\frac{k_n}{s_d}\Big)^{2k-1} \int_{S^d} F(x)^2 dx. \end{aligned}$$

Adding up $Q_2(L_nf, A_1)$ and $Q_2(L_nf, A_2)$ and using (47), we have

$$Q_2(L_nf, A_1) + Q_2(L_nf, A_2) = \Big(\frac{k_n}{s_d}\Big)^{2k-2} \frac{1}{2} \int_{(S^d)^2} (F(x)-F(y))^2 K_n^2(x,y) dxdy.$$

Now (104) follows by applying Lemma 2 to the function $(F(x)-F(y))^2$.

Now we turn to the second part of Lemma 5. We can further decompose the set A_3 into 3 subsets A_4, A_5, A_6 corresponding to $|\mathbf{T}| = 2k$ or $2k-1$ or smaller than $2k-1$. For any $(\mathbf{T},\sigma) \in A_4$, σ is neither a transposition nor identity (because $(\mathbf{T},\sigma)$

has to induce a connected graph), thus there are at least three different indices q such that $\sigma(q) \neq q$. By (41) and Lemma 3 with $r = 3$, we have

$$Q_2(L_n f, A_4) = O(n^{(2k)(d-1)} n^{-\frac{3(d-1)}{2}}) = o(n^{(d-1)(2k-1)}). \tag{107}$$

For any $(\mathbf{T}, \sigma) \in A_5$, it is not in A_2, i.e., σ is not identity, and thus there are at least two q's such that $\sigma(q) \neq q$. Applying Lemma 3 with $r = 2$, we get

$$Q_2(L_n f, A_5) = O(n^{(2k-1)(d-1)} n^{-(d-1)}) = o(n^{(d-1)(2k-1)}). \tag{108}$$

For any $(\mathbf{T}, \sigma) \in A_6$, it's clear that if $|\mathbf{T}| \leq 2k-2$, then for any σ, we have

$$|Q_2(L_n f, (\mathbf{T}, \sigma))| \leq C\Big(\frac{k_n}{s_d}\Big)^{|\mathbf{T}|} = O(n^{(d-1)|\mathbf{T}|}) = o(n^{(d-1)(2k-1)}).$$

Hence, we have

$$Q_2(L_n f, A_6) = o(n^{(d-1)(2k-1)}). \tag{109}$$

Combining (107), (108) and (109), we have

$$Q_2(L_n f, A_3) = Q_2(L_n f, A_4) + Q_2(L_n f, A_5) + Q_2(L_n f, A_6) = o(n^{(d-1)(2k-1)}),$$

which completes the proof of Lemma 5. □

Next we will prove the estimates for the higher order cumulants.

Lemma 6. *For any $m \geq 3$, it holds that*

$$Q_m(L_n f) = o(\mathrm{Var}(L_n f)^{\frac{m}{2}}), \text{ i.e., } Q_m(L_n f) = o(n^{(d-1)(km-\frac{m}{2})}). \tag{110}$$

This lemma will imply the convergence of the multivariate linear statistics to the Gaussian distribution (15) by the method of cumulants. To prove Lemma 6, we first need the following lemma.

Lemma 7. *Given a permutation σ, let $a(\sigma)$ be the number of elements q such that $\sigma(q) \neq q$. Suppose the $(\mathbf{T}, \sigma)$-graph is connected, then we have*

$$km - |\mathbf{T}| + a(\sigma) \geq m - 1 + \mathbf{1}[\sigma \neq id]. \tag{111}$$

Proof. (111) is essentially due to the simple fact in graph theory that for a connected graph the number of edges is not smaller than the number of vertices minus 1.

Before applying this fact, we note that, due to the construction of the $(\mathbf{T}, \sigma)$-graph, the connectivity property of the graph is not affected by removing some redundant red edges. Indeed, if a vertex has ℓ solid red edges, then it lies in a clique (i.e., a complete graph) of size $\ell + 1$ formed by red solid edges only. We can change this clique to a path graph by removing $\frac{\ell(\ell+1)}{2} - \ell = \frac{\ell(\ell-1)}{2}$ red solid edges without affecting the connectivity. After the edge removals, the number of solid red edges becomes $km - |\mathbf{T}|$.

We now consider the new $(\mathbf{T}, \sigma)$-graph after removing some redundant red edges as described above. Note that the total number of vertices and black edges are equal to km and $(k-1)m$, respectively.

- If $\sigma = id$, then there is no dotted red edge. The number of red solid edges (after the edge removals) is equal to $km - |\mathbf{T}|$. Hence, by the connectivity of the graph, we have

  $$(k-1)m + km - |\mathbf{T}| \geq km - 1,$$

  which proves (111).

- If $\sigma \neq id$, then we have dotted red edges. We now perform a contraction of the graph by contracting all vertices connected by black or red solid edges into a single one. After this contraction, the number of remaining vertices is at least
$$m - (km - |\mathbf{T}|).$$
These remaining vertices must be connected by dotted red edges to ensure that the $(\mathbf{T}, \sigma)$-graph is connected, whose number can be upper bounded by $a(\sigma) - 1$. (We may remove one dotted red edge without affecting the connectivity, if the number of the vertices is $a(\sigma)$.) This implies
$$a(\sigma) - 1 \geq m - (km - |\mathbf{T}|) - 1,$$
which proves (111) in the case $\sigma \neq id$.

□

Now we decompose $\mathcal{C}(m)$ into the following three subsets,
$$\begin{aligned} B_1 &= \{(\mathbf{T}, \sigma) \in \mathcal{C}(m) : |\mathbf{T}| = km, a(\sigma) = m\}, \\ B_2 &= (\mathcal{C}(m) - B_1) \cap \{(\mathbf{T}, \sigma) \in \mathcal{C}(m) : \sigma = id\}, \\ B_3 &= (\mathcal{C}(m) - B_1) \cap \{(\mathbf{T}, \sigma) \in \mathcal{C}(m) : \sigma \neq id\}. \end{aligned} \tag{112}$$
For any $(\mathbf{T}, \sigma) \in B_1$, by the restrictions that $a(\sigma) = m \geq 3$ and $(\mathbf{T}, \sigma) \in \mathcal{C}(m)$, the cycle decomposition of σ must be the multiplication of one cyclic permutation of length m and $(mk - m)$ cyclic permutations of length 1, e.g., $\sigma = (12 \cdots m)(m + 1) \cdots (km)$. Applying Lemma 4 with $r = m \geq 3$, we have
$$Q_m(L_n f, (\mathbf{T}, \sigma)) = o(n^{(d-1)|\mathbf{T}|} n^{-\frac{(d-1)m}{2}}) = o(n^{(d-1)(km - \frac{m}{2})}),$$
which further implies that
$$Q_m(L_n f, B_1) = o(n^{(d-1)(km - \frac{m}{2})}). \tag{113}$$
For any $(\mathbf{T}, \sigma) \in B_2$, by $m \geq 3$, (41), (111) and the boundedness of f, we have
$$Q_m(L_n f, (\mathbf{T}, \sigma)) = O(n^{(d-1)|\mathbf{T}|}) = O(n^{(d-1)(km - m + 1)}) = o(n^{(d-1)(km - \frac{m}{2})}).$$
Therefore, we get the estimate
$$Q_m(L_n f, B_2) = o(n^{(d-1)(km - \frac{m}{2})}). \tag{114}$$
For any $(\mathbf{T}, \sigma) \in B_3$, by the boundedness of f and Lemma 3 with $r = a(\sigma) \geq 2$, we have
$$Q_m(L_n f, (\mathbf{T}, \sigma))) = O(n^{(d-1)|\mathbf{T}|} n^{-(d-1)a(\sigma)/2})).$$
If $|\mathbf{T}| = km$, then we must have $a(\sigma) > m$ since $(\mathbf{T}, \sigma) \in \mathcal{C}(m)$ is connected but it is not in B_1. It follows that
$$O(n^{(d-1)|\mathbf{T}|} n^{-(d-1)a(\sigma)/2}) = o(n^{(d-1)(km - \frac{m}{2})}).$$
If $|\mathbf{T}| < km$, then by (111) with $\sigma \neq id$, we have
$$km - |\mathbf{T}| + \frac{a(\sigma)}{2} \geq km - |\mathbf{T}| + \frac{m - (km - |\mathbf{T}|)}{2} > \frac{m}{2},$$
which implies
$$Q_m(L_n f, (\mathbf{T}, \sigma)) = O(n^{(d-1)|\mathbf{T}|} n^{-(d-1)a(\sigma)/2}) = o(n^{(d-1)(km - \frac{m}{2})}).$$
Hence, we have
$$Q_m(L_n f, B_3) = o(n^{(d-1)(km - \frac{m}{2})}). \tag{115}$$

By (113), (114) and (115), for $m \geq 3$ we get

$$Q_m(L_n f) = Q_m(L_n f, B_1) + Q_m(L_n f, B_2) + Q_m(L_n f, B_3) = o(n^{(d-1)(km-\frac{m}{2})}).$$

This together with (105) will complete the proof of Lemma 6, and thus the proof of Theorem 2 for $k \geq 2$.

6. Proof of Theorem 3

In this section, we will prove Theorem 3. We first claim that if $f(x_1, ..., x_k)$ with $k \geq 2$ satisfies (16) and (17), then the i-margin function $f_i(x)$ is necessarily constant for all $1 \leq i \leq k$. In fact, condition (16) of the permutation invariance implies that

$$f_i(x) = f_1(x) \text{ for all } i, \tag{116}$$

which is equal to

$$\begin{aligned}
&\int_{(S^d)^{k-1}} f(x, x_2, \ldots, x_k) dx_2 \cdots dx_k \\
=&\int_{S^d} \left(\int_{(S^d)^{k-2}} f(x, x_2, \ldots, x_k) dx_3 \cdots dx_k \right) dx_2 \\
=&\int_{S^d} f_{1,2}(x, x_2) dx_2.
\end{aligned}$$

Here $f_{1,2}$ is $(1,2)$-margin function of f. Condition (17) further implies that the integral $\int_{S^d} f_{1,2}(x, x_2) dx_2$ is independent of x, i.e., $f_1(x)$ is a constant independent of x, and thus $F(x)$ is a constant. Therefore, the limit of the variance on the right hand side of (14) is degenerate. Without loss of generality, we assume that the integral of f is 0, i.e.,

$$\int_{(S^d)^k} f(x_1, \ldots, x_k) dx_1 \cdots dx_k = 0.$$

This is equivalent to $\int_{S^d} f_1(x) dx = 0$, which implies that (since f_1 is constant)

$$f_1(x) = 0 \text{ and thus } F(x) = 0 \text{ for all } x \in S^d. \tag{117}$$

6.1. **Calculations of the cumulants.** Again we will prove Theorem 3 by the method of cumulants. Recall the concepts of break points, (ir)reducible graph and the notation $\mathfrak{I}(m)$ (see Definitions 1 and 2, and (43)), we first have

Lemma 8. *Let f be a function of $k \geq 2$ variables that satisfies the i-margin function $f_i = 0$ for all i. For any $(\mathbf{T}, \sigma) \notin \mathfrak{I}(m)$, we have $Q_m(L_n f, (\mathbf{T}, \sigma)) = 0$.*

Proof. By the definition of the reducible graph, we can assume that $(\mathbf{T}, \sigma)$ breaks at $q_0 \in T_i$, and thus $\sigma(q) = q$ for $q \in \text{Range}(T_i) - q_0$. Thus we have

$$\begin{aligned}
&Q_m(L_n f, (\mathbf{T}, \sigma)) \\
=&\int_{(S^d)^{|\mathbf{T}|}} \text{sgn}(\sigma) f(T_1) \cdots f(T_m) \Pi_{q \in \text{Range}(\mathbf{T})} K_n(x_q, x_{\sigma(q)}) d\mathbf{x} \\
=&\text{sgn}(\sigma) \int_{(S^d)^{|\mathbf{T}|-k+1}} \left(\int_{(S^d)^{k-1}} f(T_i) \Pi_{q \in \text{Range}(T_i), q \neq q_0} K_n(x_q, x_q) dx_q \right) \\
&\times \left(\Pi_{j \neq i} f(T_j) \right) \left(\Pi_{q' \in \{q_0\} \cup (\text{Range}(\mathbf{T}) - \text{Range}(T_i))} K_n(x_{q'}, x_{\sigma(q')}) \right) dx_{q'} \\
=&0.
\end{aligned}$$

We have used the assumption $f_i = 0$ in the last equality. □

Lemma 8 implies that

$$Q_m(L_n f) = Q_m(L_n f, \mathcal{C}(m)) = Q_m(L_n f, \mathfrak{I}(m)).$$

Recall the concept of the circle-like graph in Definition 3, we express $\mathfrak{I}(m)$ as the union of

$$E_1 := \{(\mathbf{T}, \sigma) \in \mathfrak{I}(m) : (\mathbf{T}, \sigma) \text{ is circle-like}\} \tag{118}$$

and its complement

$$E_2 := \mathfrak{I}(m) - E_1. \tag{119}$$

Lemma 9. *For $m \geq 2$, recall that $a(\sigma)$ is the number of elements that are not fixed by σ, we have*

- *For any $(\mathbf{T}, \sigma) \in \mathfrak{I}(m)$,*

$$km - |\mathbf{T}| + \frac{a(\sigma)}{2} \geq m. \tag{120}$$

- *If $(\mathbf{T}, \sigma) \in E_2$ and $km - |\mathbf{T}| + \frac{a(\sigma)}{2} = m$, then σ is not a composition of disjoint transpositions, i.e., in the cycle decomposition of σ, there must exist at least one cyclic permutation with length strictly greater than 2.*

Proof. We now define two functions $M(i,j)$ and $\Delta(i,j)$ for $1 \leq i \leq m, 1 \leq j \leq k$. Given a $(\mathbf{T}, \sigma)$-graph, we say an index $q \in [km]$ has *multiplicity* M if there are exactly M different i's such that $q \in T_i$. We define $M(i,j)$ as the multiplicity of $T_{i,j}$. We define $\Delta(i,j) = 1$ if $\sigma(T_{i,j}) \neq T_{i,j}$ and 0 otherwise. Then we have

$$km - |\mathbf{T}| + \frac{a(\sigma)}{2} = \sum_{i=1}^{m} \sum_{j=1}^{k} \left(\frac{M(i,j) - 1}{M(i,j)} + \frac{\Delta(i,j)}{2M(i,j)} \right). \tag{121}$$

Since we assume that $(\mathbf{T}, \sigma) \in \mathfrak{I}(m)$, for each i, T_i has at least two distinct elements, denoted by T_{i,i_1} and T_{i,i_2}, such that they both have red edges. Therefore, we have

$$\max\{M(i,i_1) - 1, \Delta(i,i_1)\} \geq 1 \text{ and } \max\{M(i,i_2) - 1, \Delta(i,i_2)\} \geq 1. \tag{122}$$

If $M(i,i_1) > 1$, then

$$\frac{M(i,i_1) - 1}{M(i,i_1)} \geq \frac{1}{2}.$$

If $M(i,i_1) = 1$, then by (122), $\Delta(i,i_1) = 1$. We then have

$$\frac{\Delta(i,i_1)}{2M(i,i_1)} = \frac{\Delta(i,i_1)}{2} = \frac{1}{2}.$$

In both cases we always have

$$\frac{M(i,i_1) - 1}{M(i,i_1)} + \frac{\Delta(i,i_1)}{2M(i,i_1)} \geq \frac{1}{2}. \tag{123}$$

The same inequality holds for T_{i,i_2}. Hence we have

$$\sum_{j=1}^{k} \left(\frac{M(i,j) - 1}{M(i,j)} + \frac{\Delta(i,j)}{2M(i,j)} \right) \geq 2 \times \frac{1}{2} = 1. \tag{124}$$

And the equality in (124) holds iff there are exactly two vertices $(i, i_\alpha), \alpha \in \{1, 2\}$ that have red edges and each satisfies

$$M(i, i_\alpha) = 1 \text{ and } \Delta(i, i_\alpha) = 1; \text{ or } M(i, i_\alpha) = 2 \text{ and } \Delta(i, i_\alpha) = 0. \tag{125}$$

By summing over $1 \le i \le m$, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{k} \left(\frac{M(i,j) - 1}{M(i,j)} + \frac{\Delta(i,j)}{2M(i,j)} \right) \ge \sum_{i=1}^{m} 1 = m. \tag{126}$$

(120) now follows from (121) and (126).

Now we turn to prove the second part of Lemma 9 by contradiction. Suppose that $km - |\mathbf{T}| + a(\sigma)/2 = m$ and the cycle decomposition of σ only consists of disjoint transpositions, we need to show $(\mathbf{T}, \sigma) \in E_1$. By the proof of (120) above, the condition $km - |\mathbf{T}| + a(\sigma)/2 = m$ implies that

$$\sum_{j=1}^{k} \left(\frac{M(i,j) - 1}{M(i,j)} + \frac{\Delta(i,j)}{2M(i,j)} \right) = 1 \tag{127}$$

for each $1 \le i \le m$. This further implies that for each $1 \le i \le m$, there are exactly two vertices (i, i_1) and (i, i_2) that can have red edges, and all the other vertices have no red edges. By (125), for all $1 \le i \le m$ and any $\alpha \in \{1, 2\}$, either of the following two conditions holds:

- $M(i, i_\alpha) = 2$ and $\Delta(i, i_\alpha) = 0$. In this case (i, i_α) has exactly one solid red edge but no red dotted edge.
- $M(i, i_\alpha) = 1$ and $\Delta(i, i_\alpha) = 1$. In this case (i, i_α) has at least one dotted red edge, but no solid edge. Since σ is only composed of disjoint transpositions, (i, i_α) must have exactly one dotted red edge connecting with some other vertex $(j, j_{\alpha'})$. And j has to be distinct from i. Otherwise there would be no red edge between the set $\{(i, \cdot)\}$ and $\{(i', j') : i' \ne i, 1 \le j' \le k\}$, which makes $(\mathbf{T}, \sigma) \notin \mathcal{C}(m)$.

As a conclusion, in both cases, for each $1 \le i \le m$, there are exactly two vertices in $\{(i, j) : 1 \le j \le k\}$ that can have red edge and each of them is connected to vertices in $\{(i', j') : i' \ne i, 1 \le j' \le k\}$ with a single red edge. This shows that $(\mathbf{T}, \sigma)$ is circle-like which is a contradiction, and this proves the second part of Lemma 9. □

The following lemma indicates that the summation over the subset E_1 yields the leading order term of $Q_m(L_n f)$.

Lemma 10. *Fix any $m \ge 2$, we have the following two estimates.*

(1)

$$Q_m(L_n f, E_2) = o(n^{(d-1)(k-1)m}). \tag{128}$$

(2)

$$\begin{aligned} Q_m(L_n f, E_1) =& \frac{1}{2}(m-1)!(k(k-1))^m \left(\frac{k_n}{s_d} \right)^{mk} \left(\frac{C_d}{n^{d-1}} \right)^m \\ & \times \int_{(S^d)^m} \widehat{h}(x_1, x_2) \widehat{h}(x_2, x_3) \cdots \widehat{h}(x_m, x_1) dx_1 \cdots dx_m \\ & + o(n^{(d-1)(k-1)m}), \end{aligned} \tag{129}$$

where the constant C_d is defined in Theorem 3, and the symmetric function $\widehat{h}(x,y)$ is defined in (20).

By the relation $Q_m(L_nf) = Q_m(L_nf, \mathfrak{I}(m)) = Q_m(L_nf, E_1) + Q_m(L_nf, E_2)$, we have the following corollary.

Corollary 3. *For any $m \geq 2$, the m-th cumulant satisfies the asymptotic expansion*

$$\begin{aligned} Q_m(L_nf) =& \frac{1}{2}(m-1)!(k(k-1))^m \left(\frac{k_n}{s_d}\right)^{mk} \left(\frac{C_d}{n^{d-1}}\right)^m \\ & \times \int_{(S^d)^m} \widehat{h}(x_1,x_2)\widehat{h}(x_2,x_3)\cdots\widehat{h}(x_m,x_1)dx_1\cdots dx_m \\ & + o(n^{(d-1)(k-1)m}). \end{aligned} \tag{130}$$

In the special case $m = 2$, it yields the limit (22) *for the variance of L_nf.*

Proof of Lemma 10. We first prove part (1). Given any $(\mathbf{T},\sigma) \in E_2 \subset \mathfrak{I}(m)$, by (120), it holds that $km - |\mathbf{T}| + \frac{a(\sigma)}{2} \geq m$. For the case $km - |\mathbf{T}| + \frac{a(\sigma)}{2} > m$, by Lemma 3, we have

$$\begin{aligned} & Q_m(L_nf, (\mathbf{T},\sigma)) \\ =& \int_{(S^d)^{|\mathbf{T}|}} f(T_1)\cdots f(T_m)\mathrm{sgn}(\sigma)\Pi_{q\in \mathrm{Range}(\mathbf{T})}K_n(x_q, x_{\sigma(q)})d\mathbf{x} \\ =& O(n^{|\mathbf{T}|(d-1)})O(n^{-(d-1)a(\sigma)/2}) \\ =& O(n^{(d-1)(|\mathbf{T}|-a(\sigma)/2)}) = o(n^{(d-1)(mk-m)}). \end{aligned} \tag{131}$$

For the case $km - |\mathbf{T}| + \frac{a(\sigma)}{2} = m$, by the second part of Lemma 9, there must be a cyclic permutation whose length is at least 3 in the cycle decomposition of σ. Hence by Lemma 3 and Lemma 4, we can first integrate all variables with indices in that cyclic permutation, and then integrate the remaining variables to get

$$\begin{aligned} Q_m(L_nf, (\mathbf{T},\sigma)) &= O(n^{|\mathbf{T}|(d-1)})o(n^{-(d-1)a(\sigma)/2}) \\ &= o(n^{(d-1)(|\mathbf{T}|-a(\sigma)/2)}) = o(n^{(d-1)(mk-m)}). \end{aligned} \tag{132}$$

By (131) and (132), if we sum over all $(\mathbf{T},\sigma) \in E_2$, we prove (128).

We next prove part (2). We define

$$\begin{aligned} h_n(x,y) :=& \int_{S^d} (f_{1,2}(x,y) - f_{1,2}(x,z))P_n^2(y,z)dz \\ =& (k_n/s_d)^{-1}f_{1,2}(x,y) - \int_{S^d} f_{1,2}(x,z)P_n^2(y,z)dz. \end{aligned} \tag{133}$$

Since $f_{1,2}(x,y)$ and $P_n(x,y)$ depend only on the distance $\mathrm{d}(x,y)$, we have

$$\begin{aligned} h_n(x,y) =& (k_n/s_d)^{-1}f_{1,2}(x,y) - \int_{S^d} f_{1,2}(x,z)P_n^2(y,z)dz \\ =& (k_n/s_d)^{-1}f_{1,2}(y,x) - \int_{S^d} f_{1,2}(y,z)P_n^2(x,z)dz = h_n(y,x). \end{aligned} \tag{134}$$

Hence $h_n(x,y)$ is symmetric in x and y. We claim that

$$\begin{aligned} Q_m(L_nf,E_1) =& \frac{1}{2}(m-1)!(k(k-1))^m\left(\frac{k_n}{s_d}\right)^{mk} \\ & \times \int_{(S^d)^m} h_n(x_1,x_2)h_n(x_2,x_3)\cdots h_n(x_m,x_1)dx_1\cdots dx_m. \end{aligned} \tag{135}$$

Now we prove (135). Given $(\mathbf{T},\sigma)\in E_1$ which is circle-like, by Proposition 1, we can find vertices (i,i_1) and (i,i_2) for $1\le i\le m$ and a cyclic permutation p of $\{1,\dots,m\}$ such that (i,i_2) is connected with $(p(i),p(i)_1)$ with a red edge for all i. To compute $Q_m(L_nf,(\mathbf{T},\sigma))$, for simplicity, by condition (16) of the permutation invariance of f, we assume that $i_1=1$ and $i_2=2$ for all i, and we also assume p is the cyclic permutation $(12\cdots m)$. We now define a new kernel $\tilde{P}_i(x,y)$ as follows. If $(i,2)$ and $(i+1,1)$ are connected by a solid red edge (i.e., $T_{i,2}=T_{i+1,1}$), we let

$$\tilde{P}_i(x,y)=K_n^{-1}(x,x)\delta_y(x)=(k_n/s_d)^{-1}\delta_y(x),$$

where $\delta_y(x)$ is a Dirac delta function such that for any function g,

$$\int_{S^d}\delta_y(x)g(x)dx=g(y).$$

If $(i,2)$ and $(i+1,1)$ are connected by a dotted red edge (i.e., $\sigma(T_{i,2})=T_{i+1,1}$ or $\sigma(T_{i+1,1})=T_{i,2}$), then we let

$$\tilde{P}_i(x,y)=-P_n^2(x,y)=-(k_n/s_d)^{-2}K_n^2(x,y).$$

Integrating over all variables except those in the set $\{T_{i,\alpha},1\le i\le m,1\le\alpha\le 2\}$,

$$\begin{aligned} Q_m(L_nf,(\mathbf{T},\sigma)) = \left(\frac{k_n}{s_d}\right)^{mk}\int_{(S^d)^{2m}} & f_{1,2}(x_1,y_1)\tilde{P}_1(y_1,x_2)f_{1,2}(x_2,y_2)\tilde{P}_2(y_2,x_3)\times \\ & \cdots\times f_{1,2}(x_m,y_m)\tilde{P}_m(y_m,x_1)dx_1\cdots dx_m dy_1\cdots dy_m. \end{aligned} \tag{136}$$

If we fix the cyclic permutation $p=(1\cdots m)$ and indices $i_\alpha=1,2$, then we can get 2^m different $(\mathbf{T},\sigma)$ in the set E_1, because each red edge between $(i,2)$ and $(i+1,1)$ can either be a solid one, or a dotted one. If we sum over all 2^m different $(\mathbf{T},\sigma)$ in (136) and integrate over the variables $y_1,\dots,y_m$, then we get a total contribution of

$$\left(\frac{k_n}{s_d}\right)^{mk}\int_{(S^d)^m} h_n(x_1,x_2)h_n(x_2,x_3)\times\cdots\times h_n(x_m,x_1)dx_1\cdots dx_m. \tag{137}$$

Since there are $(m-1)!$ cyclic permutations of $[m]$ and there are $(k(k-1))^m$ distinct combinations of the indices $i_\alpha,1\le i\le m,1\le\alpha\le 2$, we obtain (135). But note that there is a factor $1/2$ in the front of (135), this is because given a circle-like $(\mathbf{T},\sigma)$-graph, the correspondence from p and $\{i_\alpha,1\le i\le m,\alpha=1,2\}$ to $(\mathbf{T},\sigma)$ is not 1-1, but rather 2-1. Indeed, by defining $p'=p^{-1}$ and $i'_\alpha=i_{3-\alpha}$, we end up at the same $(\mathbf{T},\sigma)$-graph. As an example, the $(\mathbf{T},\sigma)$-graph given in the left panel of Figure 2 is circle-like, and by Proposition 1 we can take the cyclic permutation as $p=(123)$ or $p=(132)$.

Now we prove part (2) of Lemma 10. By (79), for any fixed x and y,

$$\begin{aligned}\lim_{n\to\infty} n^{d-1}h_n(x,y) &= C_d \int_{S^d} (f_{1,2}(x,y) - f_{1,2}(x,z)) \sin^{-(d-1)}(\arccos(z\cdot y))dz \\ &= C_d \widehat{h}(x,y).\end{aligned} \tag{138}$$

Furthermore, by the boundedness of f and (52), there exists constants c and C such that for all x and y, we have

$$\left|n^{d-1}h_n(x,y)\right| \le cn^{d-1}\int_{S^d} P_n^2(y,z)dz \le C. \tag{139}$$

Hence, part (2) of Lemma 10 follows from (135), (138) and the dominated convergence theorem. □

6.2. **Identification of the limiting distribution.** Recall from (134) that h_n and thus $\widehat{h}$ are both symmetric, i.e., $\widehat{h}(x,y) = \widehat{h}(y,x)$. This implies that there exists an orthonormal basis of $L^2(S^d)$, say $w_j, j \ge 1$ such that

$$\widehat{h}(x,y) = \sum_{j=1}^{\infty} z_j w_j(x) w_j(y)$$

for almost all $(x,y) \in S^d \times S^d$.

We consider the following random variable

$$X_n := \Big(L_n f - \mathbb{E}(L_n f)\Big)\left(\frac{k_n}{s_d}\right)^{-k}\left(\frac{C_d k(k-1)}{n^{d-1}}\right)^{-1}.$$

By Corollary 3, for any fixed $m \ge 2$, we have

$$\lim_{n\to\infty} Q_m(X_n) = \frac{(m-1)!}{2}\sum_{j=1}^{\infty} z_j^m. \tag{140}$$

In addition, $Q_1(X_n) = \mathbb{E}(X_n) = 0$ for all n.

We shall now determine the specific form of the limiting distribution of X_n in three steps. Let $\chi_i, i \ge 1$ be independent chi-squared random variables with one degree of freedom, defined on some common probability space Ω_0. We consider a sequence of random variables Y_N defined by

$$Y_N = \sum_{i=1}^{N} z_i(\chi_i - 1)/2.$$

- We show that $Y_N, N \ge 1$ is a Cauchy sequence in $L^2(\Omega_0)$. Thus Y_N converges to some limiting random variable Y in the L^2 norm. We further show that the convergence is also in L^m for any $m \ge 1$, which implies that $Q_m(Y_N) \to Q_m(Y)$ for any $m \ge 1$.
- We next find the cumulants of Y by computing $\log \mathbb{E}\exp(\mathrm{i}tY_N)$ and taking the limit $N \to \infty$. It turns out that
$$Q_m(Y) = \lim_{n\to\infty} Q_m(X_n),\ \forall m \ge 1.$$
- Finally, we prove that the distribution of Y satisfies the Carleman's condition. This combined with the second step shows that X_n converges to Y in distribution and completes the proof of Theorem 3.

By (138) and (139), $\widehat{h}(x,y)$ is uniformly bounded, and thus we have

$$\sum_{j=1}^{\infty} z_j^2 = \int_{S^d}\int_{S^d} \widehat{h}(x,y)^2 dxdy < \infty.$$

Thus for any $N_1 < N_2$, it holds that

$$\|Y_{N_1} - Y_{N_2}\|_{L^2}^2 \le C \sum_{j=N_1+1}^{N_2} z_j^2 \to 0 \ \text{ as } \ N_1, N_2 \to \infty,$$

which implies that $Y_N, N \ge 1$ is a Cauchy sequence in $L^2(\Omega_0)$. Consequently, we can find a limiting random variable Y such that $Y_N \to Y$ in L^2.

For all $m \ge 2$, one has

$$\sum_{j=1}^{\infty} |z_j|^m \le \Big(\sum_{j=1}^{\infty} z_j^2\Big)^{m/2} < \infty. \tag{141}$$

By (141), for any even integer m, the m-th moment of Y_N is bounded uniformly from above:

$$\mathbb{E}(Y_N^m) \le C \sum_{m=m_1+\cdots+m_\ell : m_1,\ldots,m_\ell \ge 2} \prod_{i=1}^{\ell} \Big(\sum_{j=1}^{\infty} |z_j|^{m_i}\Big) < \infty,$$

where the summation is over all integer partitions of m. This further implies the sequence $\{Y_N^m, N \ge 1\}$ is uniformly integrable for any fixed $m \ge 1$ and thus we have

$$Y_N \xrightarrow{L^m} Y \ \text{ as } \ N \to \infty$$

for all $m \ge 1$. We can formally write Y as the sum $\sum_{i=1}^{\infty} z_i(\chi_i - 1)/2$.

We now turn to the second step. The cumulant generating function of Y_N is

$$\begin{aligned}\log \mathbb{E}\exp(\mathrm{i}tY_N) &= \sum_{i=1}^{N} \log \mathbb{E}\exp(z_i(\chi_i - 1)\mathrm{i}t/2) \\ &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{1 - z_i \mathrm{i}t}} - \sum_{i=1}^{N} \frac{z_i \mathrm{i}t}{2} = \sum_{i=1}^{N}\sum_{m=2}^{\infty} \frac{z_i^m}{2m} (\mathrm{i}t)^m.\end{aligned}$$

For $m \ge 2$, the m-th cumulant of Y_N is

$$Q_m(Y_N) = m! \sum_{j=1}^{N} \frac{z_j^m}{2m} = \frac{(m-1)!}{2} \sum_{i=1}^{N} z_i^m. \tag{142}$$

Since Y_N converges to Y in L^m for all $m \ge 1$, by (32) we have

$$Q_m(Y) = \lim_{N\to\infty} Q_m(Y_N) = \frac{(m-1)!}{2} \sum_{i=1}^{\infty} z_i^m, \ \forall\, m \ge 2, \tag{143}$$

which coincides with (140). Also, $Q_1(Y) = \mathbb{E}(Y) = 0$ since $\mathbb{E}(Y_N) = 0$ for all N.

To finish the proof of the convergence of X_n to Y, we need to show that the distribution of Y is uniquely determined by the cumulant condition (143). To this end it suffices to verify the Carleman's condition

$$\sum_{m=1}^{\infty} \big(\mathbb{E}(Y^{2m})\big)^{-1/(2m)} = \infty. \tag{144}$$

To establish (144), by (141) and (143), for $m \geq 2$, we have

$$|Q_m(Y)| \leq m! C^m, \tag{145}$$

for some constant $C > 0$. Note that (145) also holds for $m = 1$ since $Q_1(Y) = 0$. By (31) and (145), for any even integer m, we have

$$\begin{aligned} \mathbb{E}(Y^m) \leq & \sum_{R=\{R_1,\ldots,R_\ell\}\in\Pi(m)} \left|Q_{|R_1|}\right| \cdots \left|Q_{|R_\ell|}\right| \\ \leq & \sum_{R=\{R_1,\ldots,R_\ell\}\in\Pi(m)} |R_1|! C^{|R_1|} \cdots |R_\ell|! C^{|R_\ell|} \\ = & C^m \sum_{R=\{R_1,\ldots,R_\ell\}\in\Pi(m)} |R_1|! \cdots |R_\ell|!, \end{aligned} \tag{146}$$

where we used the fact that $\sum_{i=1}^{\ell} |R_i| = m$. To estimate the last summation, given an integer partition $m = m_1 + \cdots + m_\ell$ for some $\ell \geq 1$ and $m_1, \ldots, m_\ell \geq 1$. Denote the number of distinct m_i's by N' and let $\tau_1, \ldots, \tau_\ell$ be their multiplicities. Then the number of partitions $\{R_1, \ldots, R_\ell\}$ of $[m]$ such that

$$\#\{R_i : |R_i| = m_i\} = \tau_i, \quad \forall\, 1 \leq i \leq \ell$$

is given by

$$\frac{m!}{m_1! \cdots m_\ell! \prod_i^{N'} \tau_i!}.$$

Thus, we have

$$\begin{aligned} \mathbb{E}(Y^m) \leq & C^m \sum_{m=m_1+\cdots+m_\ell} \sum_{R\in\Pi(m), |R_i|=m_i} m_1! \cdots m_\ell! \\ \leq & C^m \sum_{m=m_1+\cdots+m_\ell} \frac{m!}{m_1! \cdots m_\ell! \prod_i^{N'} \tau_i!} m_1! \cdots m_\ell! \\ \leq & C^m m! \sum_{m=m_1+\cdots+m_\ell} 1. \end{aligned} \tag{147}$$

It is known that the total number of partitions of an integer m, denoted by $\kappa(m)$, satisfies $\log \kappa(m) \sim \pi\sqrt{(2m)/3}$ as $m \to \infty$ (p.70 in [1]). Consequently, for some constant $\tilde{C}$ large enough and all $m \geq 1$, $\kappa(m) \leq \tilde{C}^m$. Therefore, we have

$$\mathbb{E}(Y^m) \leq C^m m! \tilde{C}^m \leq (C\tilde{C})^m m^m. \tag{148}$$

Now (144) follows from (148) since

$$\sum_{i=1}^{\infty} \left(\mathbb{E}(Y^{2i})\right)^{-1/(2i)} \geq \sum_{i=1}^{\infty} \left((C\tilde{C})^{2i} (2i)^{2i}\right)^{-1/(2i)} \geq \sum_{i=1}^{\infty} \frac{c}{i} = \infty. \tag{149}$$

This completes the proof of (144), and thus we finish the proof of Theorem 3.

References

[1] George E. Andrews. *The theory of partitions*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1998. Reprint of the 1976 original.

[2] Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2012.

[3] B. Błaszczyszyn, D. Yogeshwaran, and J. E. Yukich. Limit theory for geometric statistics of point processes having fast decay of correlations. *Ann. Probab.*, 47(2):835–895, 2019.

[4] Omer Bobrowski and Goncalo Oliveira. Random Čech complexes on Riemannian manifolds. *Random Structures Algorithms*, 54(3):373–412, 2019.
[5] Svante Janson. *Gaussian Hilbert spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997.
[6] Matthew Kahle and Elizabeth Meckes. Limit theorems for Betti numbers of random simplicial complexes. *Homology Homotopy Appl.*, 15(1):343–374, 2013.
[7] Tomoyuki Shirai and Yoichiro Takahashi. Random point fields associated with certain Fredholm determinants. I. Fermion, Poisson and boson point processes. *J. Funct. Anal.*, 205(2):414–463, 2003.
[8] Alexander Soshnikov. The central limit theorem for local linear statistics in classical compact groups and related combinatorial identities. *Ann. Probab.*, 28(3):1353–1370, 2000.
[9] Alexander Soshnikov. Gaussian limit for determinantal random point fields. *Ann. Probab.*, 30(1):171–187, 2002.
[10] Gabor Szegő. *Orthogonal Polynomials*. American Mathematical Society Colloquium Publications, Vol. 23. American Mathematical Society, New York, 1939.
[11] D. Yogeshwaran and Robert J. Adler. On the topology of random complexes built over stationary point processes. *Ann. Appl. Probab.*, 25(6):3338–3380, 2015.

Faculty of Mathematics, Bielefeld University, Germany.
Email address: rfeng@math.uni-bielefeld.de

Faculty of Mathematics, Bielefeld University, Germany.
Email address: goetze@math.uni-bielefeld.de

Research Institute of Mathematics, Jiangsu Normal University, China.
Email address: dongyao@jsnu.edu.cn