U-statistics of determinantal point processes and Wiener chaos

Renjie Feng

Sydney Mathematical Research Institute

Oct 20th, 2024

Outline

U-statistics for i.i.d.

Determinantal point processes

Cumulants of U-statistics

Application 1: spherical ensemble and complete Wiener chaos

Application 2: infinite Ginibre ensemble and mixed χ^2 and Gaussian

U-statistics for i.i.d.

Hoeffding's form

Given i.i.d. random variables $X_1, \cdots, X_n$, Hoeffding's form for U-statistics is

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

where g is a symmetric real-valued function of k variables. WLOG, we assume $\mathbb{E}(g(X_1, .., X_k)) = 0$.

Non-degenarate case: Gaussian limit

Hoeffding (1948) proved that if $\mathrm{Var}(g(X_1, .., X_k)) < \infty$, then CLT holds

$$n^{1/2}\mathcal{U}_k(g) \xrightarrow{\mathrm{d}} N(0, k^2\delta_1).$$

Here, the constant δ_1 is the variance

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

where g_1 is the 1-margin function

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

But what if the limit is degenerate i.e., the variance $\delta_1 = 0$?

Degenerate case: χ^2-limit

If $\delta_1 = 0$, a χ^2-limit theorem holds. We suppose $g_1(x) = \mathbb{E}g(x, X_2, .., X_k) = 0$ and $\mathbb{E}g^2(X_1, .., X_k) < \infty$, then

$$n\mathcal{U}_k(g) \xrightarrow{\mathrm{d}} \binom{k}{2} \sum_{i=1}^{\infty} \lambda_i H_2(Y_i),$$

where $H_2(x) = x^2 - 1$ is the Hermite polynomial of degree 2; Y_i are i.i.d. normal distributions; λ_i are eigenvalues of the integral operator whose kernel is the symmetric 2-margin function

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

Wiener chaos decompositon

In general, $\mathcal{U}_k$ may exhibit the convergence in distribution to the Wiener chaos with arbitrary order. For example, take

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

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

$$\frac{n^{k/2}\mathcal{U}_k(g)}{\sigma^k} \xrightarrow{\mathrm{d}} H_k(Y),$$

$H_k(x)$ is the Hermite polynomial of degree k; Y is the normal distribution.

Determinantal point processes

Slater determinant

In quantum mechanics, Slater determinant describes the wave function of a multi-fermionic system.

$$\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_n(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_n(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_n) & \chi_2(\mathbf{x}_n) & \cdots & \chi_n(\mathbf{x}_n). \end{vmatrix}$$

Here, $\chi(\mathbf{x})$ is known as the spin-orbital wave function, where $\mathbf{x}$ denotes the position and spin of a single electron.

The joint density, i.e., the probability to find a particle in $\prod_i [x_i, x_i + dx_i]$, is

$$|\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)|^2 dx_1 \ldots dx_n = \det[K(\mathbf{x}_i, \mathbf{x}_j)] dx_1 \ldots dx_n,$$

where

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \chi_i(\mathbf{x}) \chi_i(\mathbf{y})$$

In general, the joint density of a determinantal point process (DPP) is

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

- Non-intersecting simple random walks
- Roots of Gaussian random function $\sum_{i=0}^{\infty} a_i z^i$
- Gaussian Unitary ensemble (Hermitian matrices + Gaussians)

$$K_n(x,y) = \left(\sum_{k=0}^{n} H_k(x) H_k(y) \right) e^{-(|x|^2+|y|^2)/2}$$

- Ginibre ensemble (square matrices + Gaussians)

$$K_n(z,w) = \left(\sum_{k=0}^{n-1} \frac{(z\bar{w})^k}{k!} \right) e^{-(|z|^2+|w|^2)/2}$$

Linear statistics

The linear statistics is

$$\mathcal{L}(f) = \sum_i f(x_i).$$

Suppose

$$\mathrm{Var}(\mathcal{L}(f)) \to \infty,$$

and

$$\exists \delta > 0,\ \mathbb{E}(\mathcal{L}(|f|)) = O(\mathrm{Var}(\mathcal{L}(f))^{\delta}),$$

then Soshinikov's CLT (2002) holds

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

U-statistics of DPPs

We consider the U-statistics of DPPs:

$$\mathcal{U}_k(g) = \sum_{X_{i_1} \neq \cdots \neq X_{i_k}} f(X_{i_1}, \ldots, X_{i_k}),$$

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

Q: Do we have Wiener chaos decomposition for U-statistics of DPPs?

A: Yes! But more complicated.

Cumulants of U-statistics

Cumulant-moment relations

Given a random variable X, its m-th cumulant $Q_m(X)$ is defined to be the coefficients in the formal expansion,

$$\log \mathbb{E} \exp(itX) = \sum_{m=1}^{\infty} \frac{Q_m(X)}{m!} (it)^m.$$

Let $\Pi(m)$ be the set of all partitions of $\{1, \ldots, m\}$, then

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

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

Method of cumulants

- First 4 terms:

$$Q_0(X) = 0, Q_1(X) = \mathbb{E}(X), Q_2(X) = \mathrm{Var}(X),$$

$$Q_3(X) = \mathbb{E}X^3 - 3\mathbb{E}X\mathbb{E}X^2 + 2(\mathbb{E}X)^3.$$

- If $X \sim N(\mu, \sigma^2)$, then $Q_m(X) = 0$ for all $m \geq 3$.
- Method of cumulants:

$$Q_m(X_n) \to Q_m(X), \ \forall\, m \geq 1 \Rightarrow X_n \to X \text{ in distribution}$$

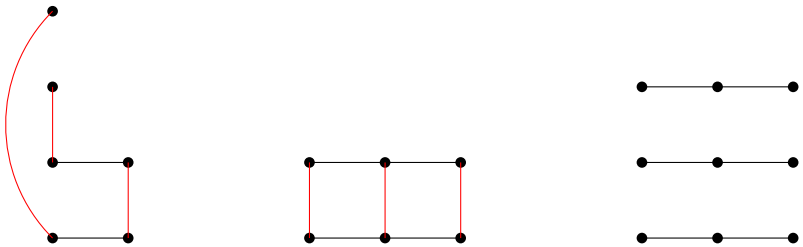
provided that X is uniquely determined by its cumulants.

The joint cumulant $Q_k(X_1, \ldots, X_k)$ is the coefficients of $i^k t_1 \cdots t_k$ in the expansion of

$$\log \mathbb{E} \exp(\sum_{j=1}^{k} i t_j X_j).$$

The joint cumulants of Hermite polynomials of central Gaussian random variables have a nice graphical representation:

$$\begin{aligned} & Q_k(H_{n_1}(X_1), \ldots, H_{n_k}(X_k)) \\ = & \sum_{\text{connected pair partitions}} \Pi_{X_i \sim X_j} \mathbb{E}(X_i X_j). \end{aligned}$$



Left: $\mathbb{E}(X_1X_4)\mathbb{E}(X_2X_3)\mathbb{E}(X_3X_4)$ in $Q_4(X_1, X_2, H_2(X_3), H_2(X_4))$
Middle: the term $(\mathbb{E}(X_1X_2))^3$ in $Q_2(H_3(X_1), H_3(X_2))$
Right: no pair partition (9 points in total, impossible to pair),
$Q_3(H_3(X_1), H_3(X_2), H_3(X_3)) = 0$.

Cumulants for linear statistics

In [3], Soshnikov derived:

$$
\begin{aligned}
& Q_m\left(\sum f(x)\right) \\
=& \sum_{\ell=1}^{m} \sum_{\bigcup_{i=1}^{\ell} V_i=[m], V_i \bigcap V_j=\emptyset, n_i=|V_i|} \int f(x_1)^{n_1} \cdots f(x_\ell)^{n_\ell} \\
& (-1)^{\ell-1} \sum_{\text{cyclic permutations } \sigma \in S_\ell} \\
& K(x_1, x_{\sigma(1)}) K(x_2, x_{\sigma(2)}) \cdots K(x_\ell, x_{\sigma(\ell)}) d\mathbf{x}.
\end{aligned}
$$

Q: Any formula for cumulants of U-statistics?

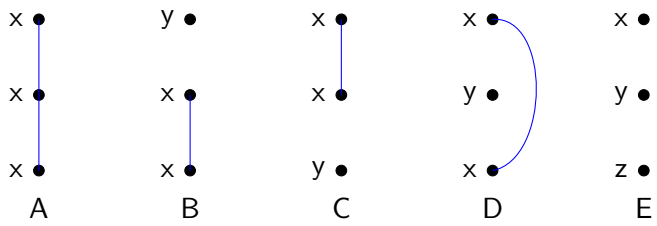
A: Yes! By graphs, which was first derived in [1] in 2022.

Reinterpretation

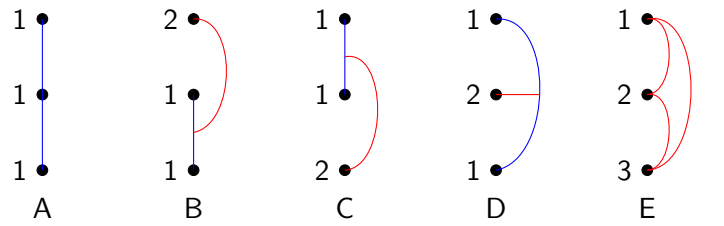
Soshnikov's formula can be reinterpreted in terms of diagram in 2 steps. For example, there are 3 terms in Q_3,

$$Q_3 = \int_{X^3} f(x)^3 K(x,x)dx - 3\int_{X^2} f(x)^2 f(y) K(x,y)K(y,x)dxdy$$
$$+2\int_{X^3} f(x)f(y)f(z)K(x,y)K(y,z)K(z,x)dxdydz.$$

We first draw **T**-graph: if two variables are the same we connect them by a blue edge. There are 5 **T**-graphs in total.



Given $\sigma \in \mathrm{Sym}(\{\mathbf{T}\})$, we can further draw an induced $(\mathbf{T}, \sigma)$-graph: we draw a red line between x and y if $\sigma(x) = y$.



$\sigma = \mathrm{id}$	(12)	(12)	(12)	(123)/(231)
$+1$	-1	-1	-1	$+1$ $+1$

The coefficients $+1$, -3 and $+2$ appear because one has to choose all $\sigma \in \mathrm{Sym}(\{\mathbf{T}\})$ such that $(\mathbf{T}, \sigma)$-graph is a connected graph.

The above 2 steps work for U-statistics. For example,

$$\mathcal{U}_3(f) = \sum_{x_1 \neq x_2 \neq x_3} f(x_1, x_2, x_3).$$

To compute the 3rd cumulant $Q_3(\mathcal{U}_3(f))$, for example $\mathbf{T} = \{T_1 = (x_1, x_2, x_3), T_2 = (x_1, x_2, x_4), T_3 = (x_5, x_6, x_7)\}$ where $|\mathbf{T}| = 7$, we draw black lines between distinct points in the same layer, and blue lines if two points in different layers are identical.

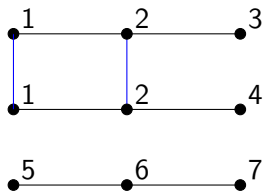


Figure: A $\mathbf{T}$-graph

Given $\sigma \in \mathrm{Sym}(\{\mathbf{T}\}) = S_7$, we can further construct a $(\mathbf{T}, \sigma)$-graph by drawing red line between x and y if $\sigma(x) = y$.

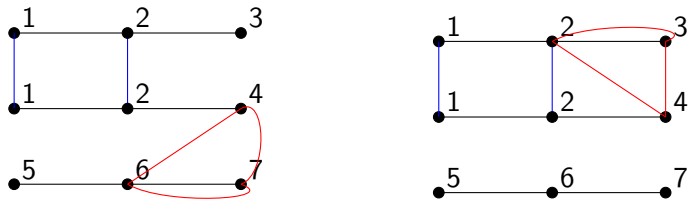


Figure: Left: the connected $(\mathbf{T}, (467))$-graph has contribution to the cumulant $Q_3(\mathcal{U}_3(f))$. Right: the disconnected $(\mathbf{T}, (234))$-graph does not contribute to cumulant.

A graphical representation formula

Define the set

$$G_m = \{(\mathbf{T}, \sigma) : (\mathbf{T}, \sigma)\text{-graph is connected.}\}.$$

Lemma 1 (F. -Yao [1], 2022)

$$Q_m(\mathcal{U}_k(f)) = \sum_{(\mathbf{T},\sigma)\in G_m} \int f(T_1)\cdots f(T_m) sgn(\sigma) \Pi_{i=1}^{q} K(x_i, x_{\sigma(i)}) d\mathbf{x}.$$

$x_1, \ldots, x_q$ *are all distinct elements in* $\mathbf{T}$.

Remark: The graph representation extends to α-DPP and Pfaffian point processes immediately.

Application 1: spherical ensemble and complete Wiener chaos

Spherical harmonics

The Laplace operator with respect to the round metric on S^d has discrete spectrum

$$\Big\{\lambda_n = -n(n+d-1), n = 0, 1, 2, ...\Big\},$$

the eigenspace

$$\mathcal{H}_n(S^d) := \mathrm{Span}_{\mathbb{R}}\{\phi : -\Delta\phi = \lambda_n\phi\}$$

and denote $d_n = \dim\mathcal{H}_n(S^d)$. Let K_n be the spectral projection

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

We consider the DPPs on S^d associated with kernel K_n.

Case 1: Gaussian limit

Define 1-margin function

$$f_1(x) = \int_{(S^d)^{k-1}} f(x, x_2, \ldots, x_k) dV(x_2) \cdots dV(x_k).$$

Let f be a bounded and symmetric function of k variables on S^d.
Assume that f_1 is not constant, then in [2], we have

$$\lim_{n\to\infty} \frac{1}{d_n^{2k-1}} \mathrm{Var}(\mathcal{U}_k(f)) = C \int_{S^d} \int_{S^d} \frac{(f_1(x) - f_1(y))^2}{\sin^{d-1}(\arccos(x \cdot y))} dV(x) dV(y).$$

In addition,

$$\frac{\mathcal{U}_k(f) - \mathbb{E}(\mathcal{U}_k(f))}{(\mathrm{Var}(\mathcal{U}_k(f)))^{\frac{1}{2}}} \xrightarrow{\mathrm{d}} N(0,1).$$

Graph and Gaussian correspondence

To prove CLT for the non-degenerate case, we need to show $Q_n(L_n f) = o(Q_2(L_n)^{m/2})$, $m \geq 3$.

- The leading term of Q_2 is given by the connected $(\mathbf{T}, \sigma)$-graphs that only have exactly one red or blue edge.

- The vanishing of the higher cumulants estimates can be derived easily.

Case 2: 2nd Wiener chaos

When the condition $f_1(x) \neq$ constant fails, $\mathcal{U}_k(f)$ can have a different limit. Define the 2-margin

$$f_2(x,y) = \int_{(S^d)^{k-2}} f(x,y,x_3 \ldots, x_k) dV(x_3)..dV(x_k).$$

If we further assume $f_2(x_1,x_2)$ depends on $\mathrm{dist}(x_1,x_2)$, then in [2]:

$$\frac{\mathcal{U}_k(f) - \mathbb{E}(\mathcal{U}_k(f))}{C' d_n^{k-1}} \xrightarrow{\mathrm{d}} \sum_{i=1}^{\infty} z_i H_2(Y_i),$$

where Y_i are i.i.d. standard Gaussian random variables, and z_i's are the eigenvalues of a Hilbert-Schmidt integral operator

$$h(x,y) := \int_{S^d} (f_2(x,y) - f_2(x,z)) \sin^{-(d-1)}(\arccos(z \cdot y)) dV(z).$$

An important class of test functions

There is an important class of test functions that satisfy these two assumptions.

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

  $\mathcal{U}_2(f)$ counts the number of pair of particles with δ distance.

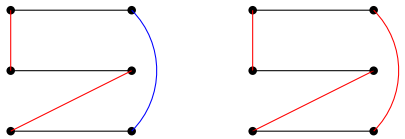
- $$f(x_1, x_2, x_3) = \mathbf{1}[d(x_1, x_2) < \delta,\ d(x_1, x_3) < \delta,\ d(x_2, x_3) < \delta]$$

  $\mathcal{U}_3(f)$ counts the number of triangles where three particles are within δ distance.

Our main result implies that both counting numbers will tend to 2nd Wiener chaos instead of Gaussian.

Higher order Wiener chaos

For example, take $f = f(x_1, x_2)$ and consider $Q_3(\mathcal{U}_2(f))$. We can show that the leading order term of $Q_3(\mathcal{U}_2(f))$ is given by the connected complete paring graphs, i.e., $(\mathbf{T}, \sigma)$-graph is connected and any point has exactly one point at other layers connecting it, for example,



Reminiscent of the pairing scheme when computing cumulants of Hermite polynomials of Gaussians!

In fact, when considering the higher order degeneration of the U-statistics, the leading order terms of cumulants of the spherical case are given by the complete paring graphs in exactly the same manner as the leading order of i.i.d. case (although no blue edge exists for i.i.d. case), which are the same graphs for cumulants of Hermite polynomials of Gaussians. Therefore, the complete Wiener chaos exist for the spherical case, albeit some different/difference operators involved in the expressions.

The similarity between the i.i.d. case and the spherical case lies on the kernel properties which imply that both variances tend to infinity. Actually this seems a general principal for other DPPs.

However, this is not the situation for the infinite Ginibre ensemble, e.g., the variance of the linear statistics is of constant.

Complete Wiener chaos

Suppose the integral operator has spectral decomposition,

$$H(x,y) = \frac{\sin^{-(d-1)}(\arccos(x \cdot y))}{\int_{S^d} \sin^{-(d-1)}(\arccos(O \cdot z))\mathrm{d}V(z)} = \sum_{i=1}^{\infty} \lambda_i u_i(x) u_i(y).$$

Note that the denominator is a constant independent of O. Then $\forall\, f$ smooth, we have the decomposition

$$f(x_1, \cdots, x_k) = \sum_{i_1,\ldots,i_k} a_{i_1,\ldots,i_k} u_{i_1}(x_1) \cdots u_{i_k}(x_k).$$

We define a transform

$$\mathrm{Op}(f) = \sum_{i_1,\ldots,i_k} \sqrt{(1-\lambda_{i_1}) \cdots (1-\lambda_{i_k})} a_{i_1,\ldots,i_k} u_{i_1}(x_1) \cdots u_{i_k}(x_k)$$

Theorem 2 (F.-Götze-Yao [2], 2023)

Let $Y_1, \dots, Y_{d_n}$ be i.i.d. Unif(S^d). Then as $n \to \infty$, the Wiener chaos expansion of $\mathcal{U}_k(f)$ is the same as

$$\tilde{\mathcal{U}}_k\left(\mathrm{Op}(f)\right) := \sum_{Y_{i_1}, \dots, Y_{i_k}} \mathrm{Op}(f)(Y_{i_1}, \dots, Y_{i_k}),$$

i.e., the U-statistics of the spherical DPPs for the test function f is the same as the U-statistics of i.i.d. uniform measure on sphere for the test function $\mathrm{Op}(f)$ in the limit.

Application 2: infinite Ginibre ensemble and mixed χ^2 and Gaussian

Infinite Ginibre ensemble

The infinite Ginibre ensemble is a determinantal point process on $\mathbb{C}$ associated with the kernel,

$$K_n(z,w) = \frac{n}{\pi} e^{nz\bar{w}} e^{-n|z|^2/2} e^{-n|w|^2/2}$$

with respect to the Lebesgue measure $d\ell$.

Let $f \in H^1 \cap L^1$. Then CLT holds for the linear statistics

$$\sum_{z \in \mathcal{Z}} f(z) - \frac{n}{\pi} \int_{\mathbb{C}} f d\ell \xrightarrow{\mathrm{d}} N(0, \frac{1}{4\pi} \|f\|^2_{H_1(\ell)}).$$

where

$$\|f\|^2_{H^1} = \int_{\mathbb{C}} |\nabla f|^2 \, d\ell, \quad \|f\|_{L^1} = \int_{\mathbb{C}} |f| d\ell.$$

For the Ginibre case, the U-statistics are more complicated (mainly because it's a slow variance case). In [2], we proved

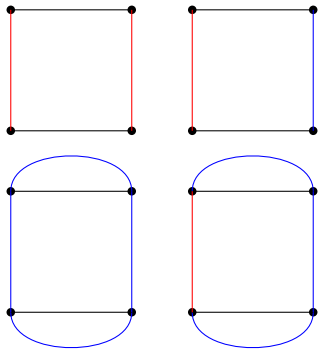
Theorem 3 (F.-Yao [1], 2022)

Let $f(x, y)$ be a symmetric smooth function on $\mathbb{C}^2$, and assume the 1-margin function

$$f_1(x) = \int f(x, y) d\ell(y) \equiv 0.$$

Then one has the limit

$$\sum_{i \neq j} f(X_i, X_j) \xrightarrow{d} \text{mixture of centered } \chi^2 + \text{ correlated } N(0, 1).$$



For example, to derive the variance, the two subfigures on the top provide the leading order term (and thus χ^2 limit) for the spherical case and i.i.d. case. In contrast, the infinite Ginibre ensemble requires two additional subfigures at the bottom.

Other examples of mixed distributions

- Subgraph count on Erdös-Rényi graph, only normal limit. But if one consider some centered U-statistics based on counting graph of k disconnected subgraphs, the limit has Wiener chaos order k. Many works around 90's.
- For subgraph count under general Graphons, the counting statistics may be a mixture of Gaussian and Chi-squared. Coefficient depending on eigenvalues of graphon integral operator. Hladký Pelekis and Šileikis (2021); Bhattacharya, Chatterjee and Janson (2023).

A remark on Pfaffian point processes

- Our graph representation holds for Pfaffian point processes similarly, where K is a self-dual quaternion kernel, e.g., circular orthogonal ensemble (COE) and circular symplectic ensemble (CSE).
- One needs to take the real part of the product of kernels by Dyson's definition of quaternion determinant.
- The problems related to the analytic aspects of Pfaffian point processes are quite open, e.g., studying U-statistics of Pfaffian point processes directly by quaternion kernels.

References

[1] R. Feng and Dong Yao, U-statistics of infinite Ginibre ensemble and Wiener chaos, 2022.

[2] R. Feng, F. Götze and D. Yao, Determinantal point processes on spheres: multivariate linear statistics. arXiv preprint arXiv: 2301.09216.

[3] A. Soshnikov. The central limit theorem for local linear statistics in classical compact groups and related combinatorial identities. Ann. Probab., 28(3):1353–1370, 2000.

[4] S. Janson. *Gaussian Hilbert spaces*, volume 129 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1997.

Thank you for your attention!