Random Matrices
MSRI Publications
Volume **65**, 2014

Three lectures on free probability

JONATHAN NOVAK

with illustrations by Michael LaCroix

These are notes from a three-lecture mini-course on free probability given at MSRI in the Fall of 2010 and repeated a year later at Harvard. The lectures were aimed at mathematicians and mathematical physicists working in combinatorics, probability, and random matrix theory. The first lecture was a staged rediscovery of free independence from first principles, the second dealt with the additive calculus of free random variables, and the third focused on random matrix models.

Introduction 310
1. Lecture one: discovering the free world 310
1.1. Counting connected graphs 310
1.2. Cumulants and connectedness 312
1.3. Cumulants and independence 315
1.4. Central limit theorem by cumulants 318
1.5. Geometrically connected graphs 319
1.6. Noncrossing cumulants 321
1.7. Noncrossing independence 328
1.8. The medium is the message 331
1.9. A brief history of the free world 333
2. Lecture two: exploring the free world 334
2.1. Random walks on the integers 334
2.2. Pólya's theorem 336
2.3. Kesten's problem 338
2.4. The classical algorithm 341
2.5. Voiculescu's algorithm 345
2.6. Solution of Kesten's problem 349
2.7. Spectral measures and free convolution 354
2.8. Free Poisson limit theorem 356
2.9. Semicircle flow 358
3. Lecture three: modelling the free world 358
3.1. Algebraic model of a free arcsine pair 358
3.2. Algebraic model of a free semicircular pair 359

3.3. Algebraic versus asymptotic models 361
3.4. Random matrix model of a free semicircular pair 364
3.5. Random matrix model of a free pair with one semicircle 370
3.6. Random matrix model of an arbitrary free pair 373
3.7. GUE + GUE 380
3.8. GUE + deterministic 380
3.9. Randomly rotated + diagonal 381
References 381

Introduction

These are notes from a three-lecture mini-course on free probability given at MSRI in the Fall of 2010 and repeated a year later at Harvard. The lectures were aimed at mathematicians and mathematical physicists working in combinatorics, probability, and random matrix theory. The first lecture was a staged rediscovery of free independence from first principles, the second dealt with the additive calculus of free random variables, and the third focused on random matrix models.

Most of my knowledge of free probability was acquired through informal conversations with my thesis supervisor, Roland Speicher, and while he is an expert in the field the same cannot be said for me. These notes reflect my own limited understanding and are no substitute for complete and rigorous treatments, such as those of Voiculescu, Dykema and Nica [Voiculescu et al. 1992], Hiai and Petz [2000], and Nica and Speicher [2006]. In addition to these sources, the expository articles of Biane [2002], Shlyakhtenko [2005] and Tao [2010] are very informative.

I would like to thank the organizers of the MSRI semester “Random Matrix Theory, Interacting Particle Systems and Integrable Systems” for the opportunity to participate as a postdoctoral fellow. Special thanks are owed to Peter Forrester for coordinating the corresponding MSRI book series volume in which these notes appear. I am also grateful to the participants of the Harvard random matrices seminar for their insightful comments and questions.

I am indebted to Michael LaCroix for making the illustrations which accompany these notes.

1. Lecture one: discovering the free world

1.1. ***Counting connected graphs.*** Let m_n denote the number of simple, undirected graphs on the vertex set $[n] = \{1, \dots, n\}$. We have $m_n = 2^{\binom{n}{2}}$, since each pair of vertices is either connected by an edge or not. A more subtle quantity is the number c_n of connected graphs on $[n]$. The sequence $(c_n)_{n\geq 1}$ is listed

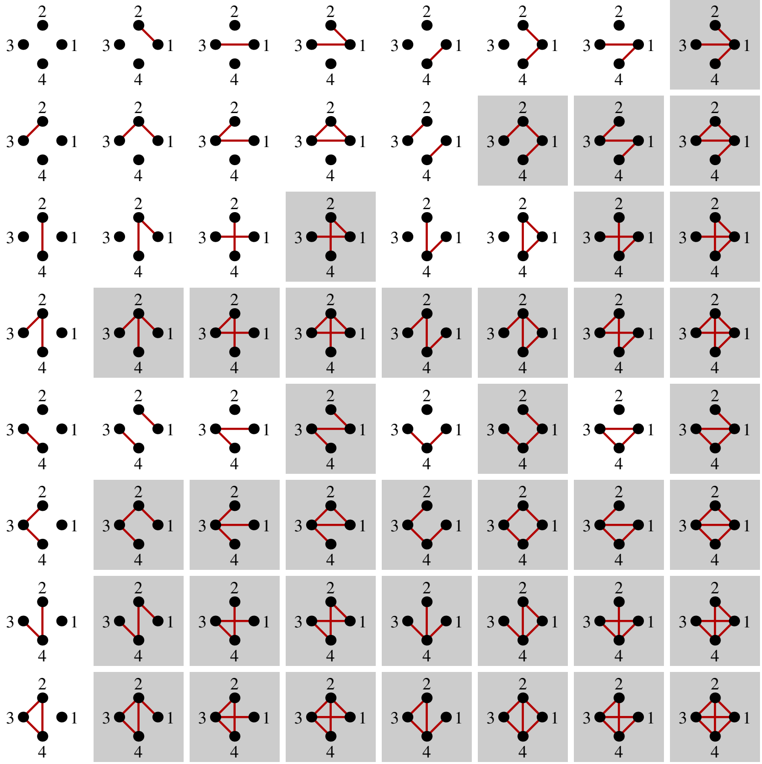

Figure 1. Thirty-eight of sixty-four graphs on four vertices are connected.

as A01187 in Sloane's Online Encyclopedia of Integer Sequences; its first few terms are

$$1,\ 1,\ 4,\ 38,\ 728,\ 26\,704,\ 1\,866\,256,\ \ldots.$$

Perhaps surprisingly, there is no closed formula for c_n. However, c_n may be understood in terms of the transparent sequence m_n in several ways, each of which corresponds to a combinatorial decomposition.

First, we may decompose a graph into two disjoint subgraphs: the connected component of a distinguished vertex, say n, and everything else, i.e., the induced subgraph on the remaining vertices. Looking at this the other way around, we may build a graph as follows. From the vertices $1, \ldots, n-1$ we can choose k of these in $\binom{n-1}{k}$ ways, and then build an arbitrary graph on these vertices in m_k ways. On the remaining $n-1-k$ vertices together with n, we may build a connected graph in c_{n-k} ways. This construction produces different graphs for different values of k, since the size of the connected component containing the pivot vertex n will be different. Moreover, as k ranges from 1 to $n-1$ we obtain all graphs in this fashion. Thus we have

$$m_n = \sum_{k=0}^{n-1} \binom{n-1}{k} m_k c_{n-k},$$

or equivalently

$$c_n = m_n - \sum_{k=1}^{n-1} \binom{n-1}{k} m_k c_{n-k}.$$

While this is not a closed formula, it allows the efficient computation of c_n given $c_1, \ldots, c_{n-1}$.

A less efficient but ultimately more useful recursion can be obtained by viewing a graph as the disjoint union of its connected components. We construct a graph by first choosing a partition of the underlying vertex set into disjoint nonempty subsets $B_1, \ldots, B_k$, and then building a connected graph on each of these, which can be done in $c_{|B_1|} \ldots c_{|B_k|}$ ways. This leads to the formula

$$m_n = \sum_{\pi \in \mathsf{P}(n)} \prod_{B \in \pi} c_{|B|},$$

where the summation is over the set of all partitions of $[n]$. We can split off the term of the sum corresponding to the partition $[n] = [n]$ to obtain the recursion

$$c_n = m_n - \sum_{\substack{\pi \in \mathsf{P}(n) \\ b(\pi) \geq 2}} \prod_{B \in \pi} c_{|B|},$$

in which we sum over partitions with at least two blocks.

The above reasoning is applicable much more generally. Suppose that m_n is the number of “structures” which can be built on a set of n labelled points, and that c_n is the number of “connected structures” on these points of the same type. Then the quantities m_n and c_n will satisfy the above (equivalent) relations. This fundamental enumerative link between connected and disconnected structures is ubiquitous in mathematics and the sciences; see [Stanley 1999, Chapter 5]. Prominent examples come from enumerative algebraic geometry [Roth 2009], where connected covers of curves are counted in terms of all covers, and quantum field theory [Etingof 2003], where Feynman diagram sums are reduced to summation over connected terms.

1.2. ***Cumulants and connectedness.*** The relationship between connected and disconnected structures is well-known to probabilists, albeit from a different point of view. In stochastic applications, $m_n = m_n(X) = \mathbb{E}[X^n]$ is the moment sequence of a random variable X, and the quantities $c_n(X)$ defined by either of the equivalent recurrences

$$m_n(X) = \sum_{k=0}^{n-1} \binom{n-1}{k} m_k(X) c_{n-k}(X)$$

and

$$m_n(X) = \sum_{\pi \in \mathsf{P}(n)} \prod_{B \in \pi} c_{|B|}(X)$$

are called the cumulants of X. This term was suggested by Harold Hotelling and subsequently popularized by Ronald Fisher and John Wishart in an influential article [1932]. Cumulants were, however, investigated as early as 1889 by the Danish mathematician and astronomer Thorvald Nicolai Thiele, who called them half-invariants. Thiele introduced the cumulant sequence as a transform of the moment sequence defined via the first of the above recurrences, and some years later arrived at the equivalent formulation using the second recurrence. The latter is now called the moment-cumulant formula. Thiele's contributions to statistics and the early theory of cumulants have been detailed by Anders Hald [1981; 2000].

Cumulants are now well-established and frequently encountered in probability and statistics, sufficiently so that the first four have been given names: mean, variance, skewness, and kurtosis.[1] The formulas for mean and variance in terms of moments are simple and familiar:

$$\begin{aligned} c_1(X) &= m_1(X), \\ c_2(X) &= m_2(X) - m_1(X)^2, \end{aligned}$$

whereas the third and fourth cumulants are more involved:

$$\begin{aligned} c_3(X) &= m_3(X) - 3m_2(X)m_1(X) + 2m_1(X)^3, \\ c_4(X) &= m_4(X) - 4m_3(X)m_1(X) - 3m_2(X)^2 + 12m_2(X)m_1(X)^2 - 6m_1(X)^4. \end{aligned}$$

It is not immediately clear why the cumulants of a random variable are of interest. If the distribution of a random variable X is uniquely determined by its moments, then we may think of the moment sequence

$$(m_1(X), m_2(X), \ldots, m_n(X), \ldots)$$

as coordinatizing the distribution of X. Passing from moments to cumulants then amounts to a (polynomial) change of coordinates. Why is this advantageous?

As a motivating example, let us compute the cumulant sequence of the most important random variable, the standard Gaussian X. The distribution of X has

[1] In practice, statisticians often define skewness and kurtosis to be the third and fourth cumulants scaled by a power of the variance.

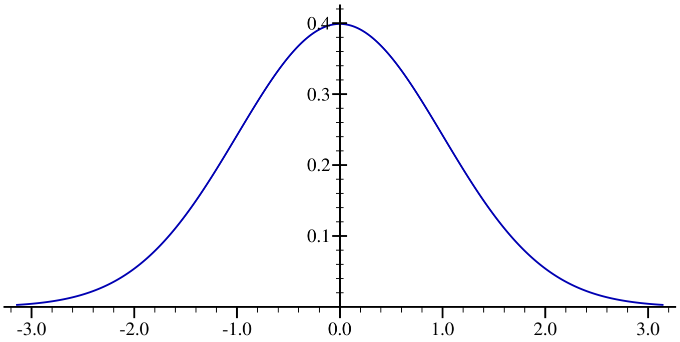

Figure 2. The Gaussian density.

density given by the bell curve

$$\mu_X(\mathrm{d}t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t,$$

depicted in Figure 2.

We will now determine the moments of X. Let z be a complex variable, and define

$$M_X(z) := \int_{\mathbb{R}} e^{tz} \mu_X(\mathrm{d}t).$$

Since $e^{-t^2/2}$ decays rapidly as $|t| \to \infty$, $M_X(z)$ is a well-defined entire function of z whose derivatives can be computed by differentiation under the integral sign:

$$M_X'(z) = \int_{\mathbb{R}} t e^{tz} \mu_X(\mathrm{d}t), \quad M_X''(z) = \int_{\mathbb{R}} t^2 e^{tz} \mu_X(\mathrm{d}t), \quad \ldots .$$

In particular, the n-th derivative of $M_X(z)$ at $z = 0$ is

$$M_X^{(n)}(0) = \int_{\mathbb{R}} t^n \mu_X(\mathrm{d}t) = m_n(X),$$

so we have the Maclaurin series expansion

$$M_X(z) = \sum_{n=0}^{\infty} m_n(X) \frac{z^n}{n!}.$$

Thus, the integral $M_X(z)$ acts as an exponential generating function for the moments of X. On the other hand, this integral may be explicitly evaluated. Completing the square in the exponent of the integrand we find that

$$M_X(z) = e^{z^2/2} \int_{\mathbb{R}} e^{-(t-z)^2/2} \frac{\mathrm{d}t}{\sqrt{2\pi}},$$

whence

$$M_X(z) = e^{z^2/2} = \sum_{k=0}^{\infty} \frac{z^{2k}}{2^k k!}$$

for real z by translation invariance of Lebesgue measure, and hence for all $z \in \mathbb{C}$. We conclude that the odd moments of X vanish, while the even ones are given by the formula

$$m_{2k}(X) = \frac{(2k)!}{2^k k!} = (2k-1) \cdot (2k-3) \cdot \cdots \cdot 5 \cdot 3 \cdot 1.$$

This is the number of partitions of the set $[2k]$ into blocks of size two, also called "pairings": we have $2k-1$ choices for the element to be paired with 1, then $2k-3$ choices for the element to be paired with the smallest remaining unpaired element, etc. Alternatively, we may say that $m_n(X)$ is equal to the number of 1-regular graphs on n labelled vertices. It now follows from the fundamental link between connected and disconnected structures that the cumulant $c_n(X)$ is equal to the number of connected 1-regular graphs. Consequently, the cumulant sequence of a standard Gaussian random variable is simply

$$(0, 1, 0, 0, 0, \dots)$$

That the universality of the Gaussian distribution is reflected in the simplicity of its cumulant sequence signals cumulants as a key concept in probability theory. In Thiele's own words [Hald 2000],

> This remarkable proposition has originally led me to prefer the half-invariants over every other system of symmetrical functions.

This sentiment persists amongst modern-day probabilists. To quote Terry Speed [1983],

> In a sense which it is hard to make precise, all of the important aspects of distributions seem to be simpler functions of cumulants than of anything else, and they are also the natural tools with which transformations of systems of random variables can be studied when exact distribution theory is out of the question.

1.3. ***Cumulants and independence.*** The importance of cumulants stems, ultimately, from their relationship with stochastic independence. Suppose that X and Y are a pair of independent random variables whose moment sequences have been given to us by an oracle, and our task is to compute the moments of $X+Y$. Since $\mathbb{E}[X^a Y^b] = \mathbb{E}[X^a]\,\mathbb{E}[Y^b]$, this can be done using the formula

$$m_n(X+Y) = \sum_{k=0}^{n} \binom{n}{k} m_k(X) m_{n-k}(Y),$$

which is conceptually clear but computationally inefficient because of its dependence on n. For example, if we want to compute $m_{100}(X+Y)$ we must evaluate a sum with 101 terms, each of which is a product of three factors. Computations with independent random variables simplify dramatically if one works with cumulants rather than moments. Indeed, Thiele called cumulants "half-invariants" because

$$X, Y \text{ independent } \implies c_n(X+Y) = c_n(X) + c_n(Y) \;\; \forall n \geq 1.$$

Thanks to this formula, if the cumulant sequences of X and Y are given, then each cumulant of $X+Y$ can be computed simply by adding two numbers. The mantra to be remembered is that

$$\boxed{\textit{cumulants linearize addition of independent random variables}.}$$

For example, this fact together with the computation we did above yields that the sum of two iid standard Gaussians is a Gaussian of variance two.

In order to precisely understand the relationship between cumulants and independence, we need to extend the relationship between moments and cumulants to a relationship between mixed moments and mixed cumulants. Mixed moments are easy to define: given a set of (not necessarily distinct) random variables $X_1, \dots, X_n$,

$$m_n(X_1, \dots, X_n) := \mathbb{E}[X_1 \dots X_n].$$

It is clear that $m_n(X_1, \dots, X_n)$ is a symmetric, multilinear function of its arguments. The new notation for mixed moments is related to our old notation for pure moments by

$$m_n(X) = m_n(X, \dots, X),$$

which we may keep as a useful shorthand.

We now define mixed cumulants recursively in terms of mixed moments using the natural extension of the moment-cumulant formula:

$$m_n(X_1, \dots, X_n) = \sum_{\pi \in \mathsf{P}(n)} \prod_{B \in \pi} c_{|B|}(X_i : i \in B).$$

For example, we have

$$m_2(X_1, X_2) = c_2(X_1, X_2) + c_1(X_1)c_1(X_2),$$

from which we find that the second mixed cumulant of X_1 and X_2 is their covariance,

$$c_2(X_1, X_2) = m_2(X_1, X_2) - m_1(X_1)m_2(X_2).$$

More generally, the recurrence

$$c_n(X_1, \dots, X_n) = m_n(X_1, \dots, X_n) - \sum_{\substack{\pi \in \mathsf{P}(n) \\ b(\pi) \geq 2}} \prod_{B \in \pi} c_{|B|}(X_i : i \in B)$$

facilitates a straightforward inductive proof that $c_n(X_1, \dots, X_n)$ is a symmetric, n-linear function of its arguments, which explains Thiele's reference to cumulants as his preferred system of symmetric functions.

The fundamental relationship between cumulants and stochastic independence is the following: X and Y are independent if and only if all their mixed cumulants vanish:

$$\begin{aligned}
&c_2(X, Y) = 0, \\
&c_3(X, X, Y) = c_3(X, Y, Y) = 0, \\
&c_4(X, X, X, Y) = c_4(X, X, Y, Y) = c_4(X, Y, Y, Y) = 0, \\
&\qquad \vdots
\end{aligned}$$

The forward direction of this theorem,

$$X, Y \text{ independent} \implies \text{mixed cumulants vanish},$$

immediately yields Thiele's linearization property, since by multilinearity we have

$$\begin{aligned}
c_n(X + Y) &= c_n(X + Y, \dots, X + Y) \\
&= c_n(X, \dots, X) + \text{ mixed cumulants } + c_n(Y, \dots, Y) \\
&= c_n(X) + c_n(Y).
\end{aligned}$$

Conversely, let X, Y be a pair of random variables whose mixed cumulants vanish. Let us check in a couple of concrete cases that this condition forces X and Y to obey the algebraic identities associated with independent random variables. In the first nontrivial case, $n = 2$, vanishing of mixed cumulants reduces the extended moment-cumulant formula to

$$m_2(X, Y) = c_1(X)c_1(Y) = m_1(X)m_1(Y),$$

which is consistent with the factorization rule $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ for independent random variables. Now let us try an $n = 4$ example. We compute $m_4(X, X, Y, Y)$ directly from the extended moment cumulant formula. Referring to Figure 3, we find that vanishing of mixed cumulants implies

$$\begin{aligned}
m_4(X, X, Y, Y) = c_2(X, X)c_2(Y, Y) + c_2(X, X)c_1(Y)c_1(Y) \\
+ c_2(Y, Y)c_1(X)c_1(X) + c_1(X)c_1(X)c_1(Y)c_1(Y),
\end{aligned}$$

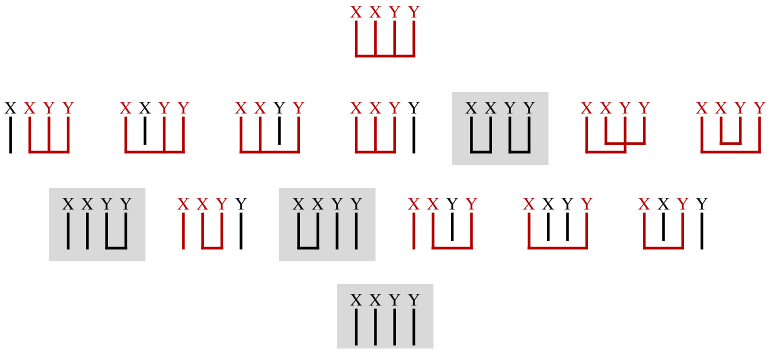

Figure 3. Graphical evaluation of $m_4(X, X, Y, Y)$.

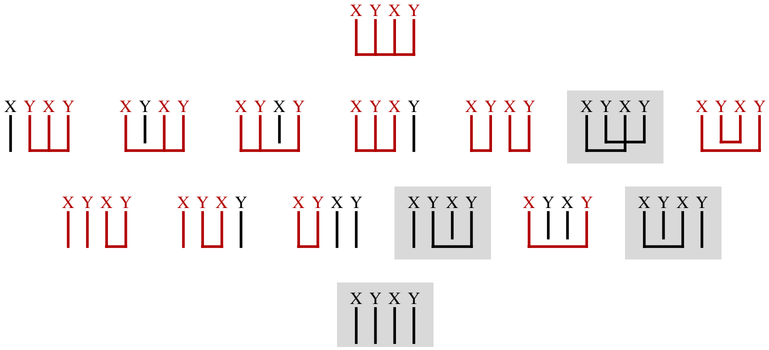

Figure 4. Graphical evaluation of $m_4(X, Y, X, Y)$.

which reduces to the factorization identity $\mathbb{E}[X^2Y^2] = \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$.

Of course, if we compute $m_4(X, Y, X, Y)$ using the extended moment-cumulant formula we should get the same answer, and indeed this is the case, but it is important to note that the contributions to the sum come from different partitions, as indicated in Figure 4.

1.4. ***Central limit theorem by cumulants.*** We can use the theory of cumulants presented thus far to prove an elementary version of the central limit theorem. Let $X_1, X_2, X_3 \dots$ be a sequence of iid random variables, and let X be a standard Gaussian. Suppose that the common distribution of the variables X_i has mean zero, variance one, and finite moments of all orders. Put

$$S_N := \frac{X_1 + \cdots + X_N}{\sqrt{N}}.$$

Then, for each positive integer n,

$$\lim_{N\to\infty} m_n(S_N) = m_n(X).$$

Since moments and cumulants mutually determine one another, in order to prove this CLT it suffices to prove that

$$\lim_{N\to\infty} c_n(S_N) = c_n(X)$$

for each $n \geq 1$. Now, by the multilinearity of c_n and the independence of the X_i, we have

$$\begin{aligned} c_n(S_N) &= c_n(N^{-1/2}(X_1 + \cdots + X_N)) \\ &= N^{-n/2}(c_n(X_1) + \cdots + c_n(X_N)) \\ &= N^{1-n/2} c_n(X_1), \end{aligned}$$

where the last line follows from the fact that the X_i are equidistributed. Thus: if $n = 1$,

$$c_1(S_N) = N^{1/2} c_1(X_1) = 0;$$

if $n = 2$,

$$c_2(S_N) = c_2(X_1) = 1;$$

if $n > 2$,

$$c_n(S_N) = N^{\text{negative number}} c_n(X_1).$$

We conclude that

$$\lim_{N\to\infty} c_n(S_N) = \delta_{n2},$$

which we have already identified as the cumulant sequence of a standard Gaussian random variable.

1.5. ***Geometrically connected graphs.*** Let us now consider a variation on our original graph-counting question. Given a graph G on the vertex set $[n]$, we may represent its vertices by n distinct points on the unit circle (say, the n-th roots of unity) and its edges by straight line segments joining these points. This is how we represented the set of four-vertex graphs in Figure 1. We will denote this geometric realization of G by $|G|$. The geometric realization of a graph carries extra structure which we may wish to consider. For example, it may happen that $|G|$ is a connected set of points in the plane even if the graph G is not connected in the usual sense of graph theory. Let κ_n denote the number of geometrically connected graphs on $[n]$. This is sequence A136653 in Sloane's database; its first few terms are

$$1,\ 1,\ 4,\ 39,\ 748,\ 27\,162,\ 1\,880\,872,\ \ldots.$$

Since geometric connectivity is a weaker condition than set-theoretic connectivity, κ_n grows faster than c_n; these sequences diverge from one another at $n = 4$, where

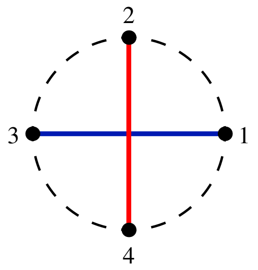

Figure 5. The crosshairs graph.

the unique disconnected but geometrically connected graph is the "crosshairs" graph shown in Figure 5.

Consider now the problem of computing κ_n. As with c_n, we can address this problem by means of a combinatorial decomposition of the set of graphs with n vertices. However, this decomposition must take into account the planar nature of geometric connectivity, which our previous set-theoretic decompositions do not. Consequently, we must formulate a new decomposition.

Given a graph G on $[n]$, let $\pi(G)$ denote the partition of $[n]$ induced by the connected components of G (i and j are in the same block of $\pi(G)$ if and only if they are in the same connected component of G), and let $\pi(|G|)$ denote the partition of $[n]$ induced by the geometrically connected components of $|G|$ (i and j are in the same block of $\pi(|G|)$ if and only if they are in the same geometrically connected component of $|G|$). How are $\pi(G)$ and $\pi(|G|)$ related? To understand this, let us view our geometric graph realizations as living in the hyperbolic plane rather than the Euclidean plane. Thus Figure 1 depicts line systems in the Klein model, in which the plane is an open disc and straight lines are chords of the boundary circle. We could alternatively represent a graph in the Poincaré disc model, where straight lines are arcs of circles orthogonal to the boundary circle, or in the Poincaré half-plane model, where space is an open-half plane and straight lines are arcs of circles orthogonal to the boundary line. The notion of geometric connectedness does not depend on the particular realization chosen. The half-plane model has the useful feature that the geometric realization $|G|$ essentially coincides with the pictorial representation of $\pi(G)$, and we can see clearly that crossings in $|G|$ correspond exactly to crossings in $\pi(G)$. Thus, $\pi(|G|)$ is obtained by fusing together crossing blocks of $\pi(G)$. The resulting partition $\pi(|G|)$ no longer has any crossings — by construction, it is a noncrossing partition; see Figure 6.

We can now obtain a recurrence for κ_n. We construct a graph by first choosing a noncrossing partition of the underlying vertex set into blocks $B_1, \dots, B_k$ and then building a geometrically connected graph on each block, which can be done

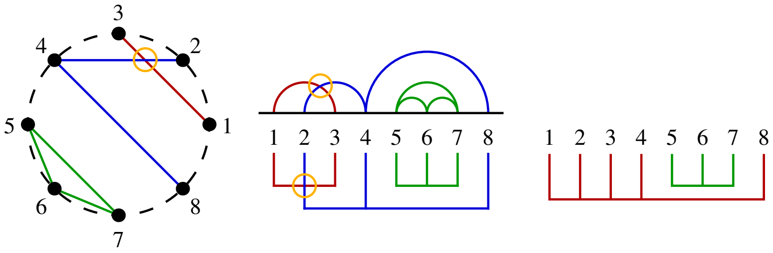

Figure 6. Partition fusion accounts for geometric connectedness.

in $\kappa_{|B_1|} \dots \kappa_{|B_k|}$ ways. This leads to the formula

$$m_n = \sum_{\pi \in \mathsf{NC}(n)} \prod_{B \in \pi} \kappa_{|B|},$$

where the summation is over noncrossing partitions of $[n]$. Just as before, we can split off the term of the sum corresponding to the partition with only one block to obtain the recursion

$$\kappa_n = m_n - \sum_{\substack{\pi \in \mathsf{NC}(n) \\ b(\pi) \geq 2}} \prod_{B \in \pi} \kappa_{|B|},$$

in which we sum over noncrossing partitions with at least two blocks.

1.6. ***Noncrossing cumulants.*** We have seen above that the usual graph theoretic notion of connectedness manifests itself probabilistically as the cumulant concept. We have also seen that graph theoretic connectedness has an interesting geometric variation, which we called geometric connectedness. This begs the question:

Is there a probabilistic interpretation of geometric connectedness?

Let X be a random variable, with moments $m_n(X)$. Just as the classical cumulants $c_n(X)$ were defined recursively using the relation between all structures and connected structures, we define the noncrossing cumulants of X recursively using the relation between all structures and geometrically connected structures:

$$m_n(X) = \sum_{\mathsf{NC}(n)} \prod_{B \in \pi} \kappa_{|B|}(X).$$

We will call this the noncrossing moment-cumulant formula. Since connectedness and geometric connectedness coincide for structures of size $n = 1, 2, 3$, the first three noncrossing cumulants of X are identical to its first three classical cumulants. However, for $n \geq 4$, the noncrossing cumulants become genuinely new statistics of X.

Our first step in investigating these new statistics is to look for a noncrossing analogue of the most important random variable, the standard Gaussian. This should be a random variable whose noncrossing cumulant sequence is

$$0,\ 1,\ 0,\ 0,\ \dots .$$

If this search leads to something interesting, we may be motivated to further investigate noncrossing probability theory.

From the noncrossing moment-cumulant formula, we find that the moments of the noncrossing Gaussian X are given by

$$m_n(X) = \sum_{\pi \in \mathsf{NC}(n)} \prod_{B \in \pi} \delta_{|B|,2} = \sum_{\pi \in \mathsf{NC}_2(n)} 1.$$

That is, $m_n(X)$ is equal to the number of partitions in $\mathsf{NC}(n)$ all of whose blocks have size 2, i.e., noncrossing pairings of n points. We know that there are no pairings at all on an odd number of points, so the odd moments of X must be zero, which indicates that X likely has a symmetric distribution. The number of pairings on $n = 2k$ points is given by a factorial going down in steps of two, $(2k-1)!! = (2k-1) \cdot (2k-3) \cdot \dots \cdot 5 \cdot 3 \cdot 1$, so the number of noncrossing pairings must be smaller than this double factorial.

In order to count noncrossing pairings on $2k$ points, we construct a function f from the set of all pairings on $2k$ points to length $2k$ sequences of ± 1. This function is easy to describe: if $i < j$ constitute a block of π, then the i-th element of $f(\pi)$ is $+1$ and the j-th element of $f(\pi)$ is -1. See Figure 7 for an illustration of this function in the case $k = 3$. By construction, f is a surjection from the set of pairings on $2k$ points onto the set of length $2k$ sequences of ± 1 all of whose partial sums are nonnegative and whose total sum is zero. We leave it to the reader to show that the fibre of f over any such sequence contains exactly one noncrossing pairing, so that f restricts to a bijection from noncrossing pairings onto its image. The image sequences can be neatly enumerated using the Dvoretzky–Motzkin–Raney cyclic shift lemma, as in [Graham et al. 1989, Section 7.5]. They are counted by the Catalan numbers

$$\mathrm{Cat}_k = \frac{1}{k+1}\binom{2k}{k},$$

which are smaller than the double factorials by a factor of $2^k/(k+1)!$. In fact, since $\mathrm{Cat}_k < 4^k$, we can conclude that the distribution of X is compactly supported.

We have discovered that

$$m_n(X) = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \mathrm{Cat}_{n/2} & \text{if } n \text{ even.} \end{cases}$$

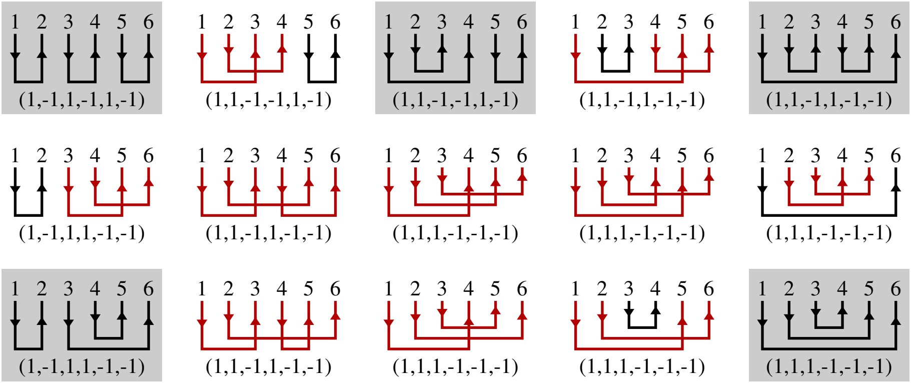

Figure 7. Construction of the function f from pairings to bitstrings.

The Catalan numbers are ubiquitous in enumerative combinatorics (see [Stanley 1999, Exercise 6.19] as well as [Stanley 2013]), and their appearance in this context is the first sign that we are onto something interesting. We are now faced with an inverse problem: we are not trying to calculate the moments of a random variable given its distribution, rather we know that the moment sequence of X is

$$0,\ \mathrm{Cat}_1,\ 0,\ \mathrm{Cat}_2,\ 0,\ \mathrm{Cat}_3,\ 0,\ \ldots.$$

and we would like to write down its distribution μ_X. Equivalently, we are looking for an integral representation of the entire function

$$M_X(z) = \sum_{n=0}^{\infty} \mathrm{Cat}_n \frac{z^{2n}}{(2n)!} = \sum_{n=0}^{\infty} \frac{z^{2n}}{n!(n+1)!}$$

which has the form

$$M_X(z) = \int_{\mathbb{R}} e^{tz} \mu_X(\mathrm{d}t),$$

with μ_X a probability measure on the real line. The solution to this problem can be extracted from the classical theory of Bessel functions.

The modified Bessel function $I_\alpha(z)$ of order α is one of two linearly independent solutions to the modified Bessel equation

$$\left(z^2 \frac{d^2}{dz^2} + z\frac{d}{dz} - (z^2+\alpha^2)\right) F = 0,$$

the other being the Macdonald function

$$K_\alpha(z) = \frac{\pi}{2} \frac{I_{-\alpha}(z) - I_\alpha(z)}{\sin(\alpha\pi)}.$$

The modified Bessel equation (and hence the functions I_α, K_α) appears in many problems of physics and engineering since it is related to solutions of Laplace's equation with cylindrical symmetry. An excellent reference on this topic is [Andrews et al. 1999, Chapter 4].

Interestingly, Bessel functions also occur in the combinatorics of permutations: a remarkable identity due to Ira Gessel asserts that

$$\det[I_{i-j}(2z)]_{i,j=1}^{k}=\sum_{n=0}^{\infty}\operatorname{lis}_k(n)\frac{z^{2n}}{(n!)^2},$$

where $\operatorname{lis}_k(n)$ is the number of permutations in the symmetric group $\mathbf{S}(n)$ with no increasing subsequence of length $k+1$. Gessel's identity was the point of departure in the work of Jinho Baik, Percy Deift and Kurt Johansson who, answering a question posed by Stanislaw Ulam, proved that the limit distribution of the length of the longest increasing subsequence in a uniformly distributed random permutation is given by the ($\beta=2$) Tracy–Widom distribution. This nonclassical distribution was isolated and studied by Craig Tracy and Harold Widom in a series of works on random matrix theory in the early 1990s where it emerged as the limiting distribution of the top eigenvalue of large random Hermitian matrices. It has a density which may also be described in terms of Bessel functions, albeit indirectly. Consider the ordinary differential equation

$$\frac{d^2}{dx^2}u=2u^3+xu$$

for a real function $u=u(x)$, which is known as the Painlevé II equation after the French mathematician (and two-time Prime Minister of France) Paul Painlevé. It is known that this equation has a unique solution, called the Hastings–McLeod solution, with the asymptotics $u(x)\sim-\operatorname{Ai}(x)$ as $x\to\infty$, where

$$\operatorname{Ai}(x)=\frac{1}{\pi}\sqrt{\frac{x}{3}}K_{\frac{1}{3}}(\tfrac{2}{3}x^{3/2})$$

is a scaled specialization of the Macdonald function known as the Airy function. Define the Tracy–Widom distribution function by

$$F(t)=e^{-\int_t^\infty(x-t)u(x)^2\mathrm{d}x},$$

where u is the Hastings–McLeod solution to Painlevé II. The theorem of Baik, Deift and Johansson asserts that

$$\lim_{n\to\infty}\frac{1}{n!}\operatorname{lis}_{2\sqrt{n}+tn^{1/6}}(n)=F(t)$$

for any $t\in\mathbb{R}$. From this one may conclude, for example, that the probability a permutation drawn uniformly at random from the symmetric group $\mathbf{S}(n^2)$ avoids the pattern 1 2 ... $2n+1$ converges to $F(0) = 0.9694\ldots$. We refer the interested reader to Richard Stanley's survey [2007] for more information on this topic.

Nineteenth century mathematicians knew how to describe the modified Bessel function both as a series,

$$I_\alpha(z) = \sum_{n=0}^{\infty} \frac{(\frac{z}{2})^{2n+\alpha}}{n!\,\Gamma(n+1+\alpha)},$$

and as an integral,

$$I_\alpha(z) = \frac{(\frac{z}{2})^{\alpha}}{\sqrt{\pi}\,\Gamma(\alpha+\frac{1}{2})} \int_0^{\pi} e^{(\cos\theta)z} (\sin\theta)^{2\alpha}\, \mathrm{d}\theta.$$

From the series representation we find that

$$M_X(z) = \frac{I_1(2z)}{z},$$

and consequently we have the integral representation

$$M_X(z) = \frac{2}{\pi} \int_0^{\pi} e^{2(\cos\theta)z} \sin^2\theta\, \mathrm{d}\theta.$$

This is one step removed from what we want: it tells us that the Catalan numbers are the even moments of the random variable $X = 2\cos(Y)$, where Y is a random variable with distribution

$$\mu_Y(\mathrm{d}\theta) = \frac{2}{\pi} \sin^2\theta\, \mathrm{d}\theta$$

supported on the interval $[0, \pi]$. However, this is a rather interesting intermediate step since the above measure appears in number theory, where it is called the Sato–Tate distribution; see Figure 8.

The Sato–Tate distribution arises in the arithmetic statistics of elliptic curves. The location of integer points on elliptic curves is a classical topic in number

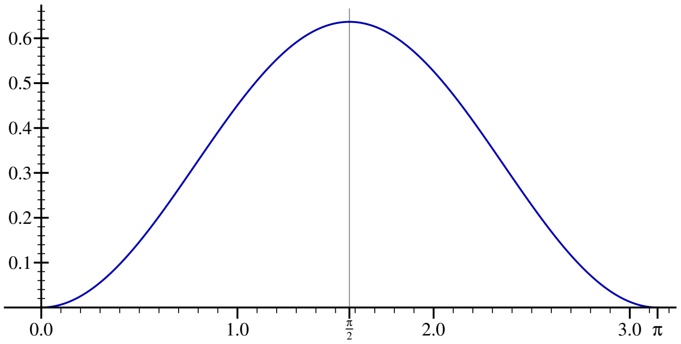

Figure 8. The Sato–Tate density.

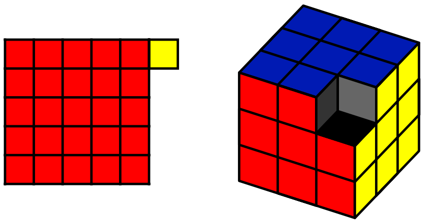

Figure 9. Diophantine perspectives on twenty-six.

theory. For example, Diophantus of Alexandria wrote that the equation

$$y^2 = x^3 - 2$$

has the solution $x = 3$, $y = 5$, and in the 1650s Pierre de Fermat claimed that there are no other positive integer solutions. This is the striking assertion that 26 is the only number one greater than a perfect square and one less than a perfect cube (see Figure 9). That this is indeed the case was proved by Leonhard Euler in 1770, although according to some sources Euler's proof was incomplete and the solution to this problem should be attributed to Axel Thue in 1908.

Modern number theorists study solutions to elliptic Diophantine equations by reducing modulo primes. Given an elliptic curve

$$y^2 = x^3 + ax + b, \qquad a, b \in \mathbb{Z},$$

let $\Delta = -16(4a^3 + 27b^2)$ be sixteen times the discriminant of $x^3 + ax + b$, and let S_p be the number of solutions of the congruence

$$y^2 \equiv x^3 + ax + b \qquad \text{mod } p$$

where p is a prime which does not divide Δ. In his 1924 doctoral thesis, Emil Artin conjectured that

$$|S_p - p| \leq 2\sqrt{p}$$

for all such good reduction primes. This remarkable inequality states that the number of solutions modulo p is roughly p itself, up to an error of order $\sqrt{p}$. Artin's conjecture was proved by Helmut Hasse in 1933. Around 1960, Mikio Sato and John Tate became interested in the finer question of the distribution of the centred and scaled solution count $(S_p - p)/\sqrt{p}$ for typical elliptic curves E (meaning those without complex multiplication) as p ranges over the infinitely many primes not dividing the discriminant of E. Because of Hasse's theorem, this amounts to studying the distribution of the angle θ_p defined by

$$\frac{S_p - p}{\sqrt{p}} = 2\cos\theta_p$$

in the interval $[0, \pi]$. Define a sequence μ_N^E of empirical probability measures associated to E by

$$\mu_N^E = \frac{1}{\pi(N)} \sum_{p \leq N} \delta_{\theta_p},$$

where $\pi(N)$ is the number of prime numbers less than or equal to N. Sato and Tate conjectured that, for any elliptic curve E without complex multiplication, μ_N^E converges weakly to the Sato–Tate distribution as $N \to \infty$. This is a universality conjecture: it posits that certain limiting behaviour is common to a large class of elliptic curves irrespective of their fine structural details. Major progress on the Sato–Tate conjecture has been made within the last decade; we refer the reader to the surveys of Barry Mazur [2006] and Ram Murty and Kumar Murty [2009] for further information.

The random variable we seek is not the Sato–Tate variable Y, but twice its cosine, $X = 2\cos Y$. Making the substitution $s = \arccos\theta$ in the integral representation of $M_X(z)$ obtained above, we obtain

$$M_X(z) = \frac{2}{\pi} \int_{-1}^{1} e^{2sz} \sqrt{1-s^2}\, \mathrm{d}s,$$

and further substituting $t = 2s$ this becomes

$$M_X(z) = \frac{1}{2\pi} \int_{-2}^{2} e^{tz} \sqrt{4-t^2}\, \mathrm{d}t.$$

Thus the random variable X with even moments the Catalan numbers and vanishing odd moments

$$\mu_X(\mathrm{d}t) = \frac{1}{2\pi} \sqrt{4-t^2}\, \mathrm{d}t,$$

which is both symmetric and compactly supported. This is another famous distribution: it is called the Wigner semicircle distribution after the physicist Eugene Wigner, who considered it in the 1950s in a context ostensibly unrelated to elliptic curves. The density of μ_X is shown in Figure 10 — note that it is not a semicircle, but rather half an ellipse of semi-major axis two and semi-minor axis $1/\pi$.

Wigner was interested in constructing models for the energy levels of complex systems, and hit on the idea that the eigenvalues of large symmetric random matrices provide a good approximation. Wigner considered $N \times N$ symmetric matrices X_N whose entries $X_N(ij)$ are independent random variables, up to the symmetry constraint $X_N(ij) = X_N(ji)$. Random matrices of this form are now known as Wigner matrices, and their study remains a topic of major interest today. Wigner studied the empirical spectral distribution of the eigenvalues of

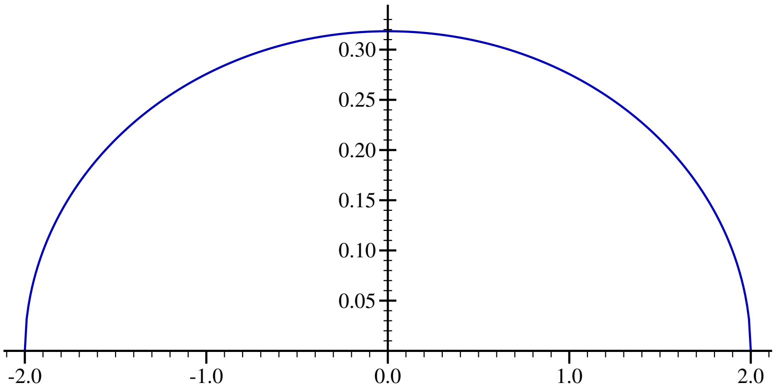

Figure 10. The Wigner semicircle density.

X_N, i.e., the probability measure

$$\mu_N = \frac{1}{N} \sum_{k=1}^{N} \delta_{\lambda_k(N)}$$

which places mass $1/N$ at each eigenvalue of X_N. Note that, unlike in the setting above where we considered the sequence of empirical measures associated to a fixed elliptic curve E, the measure μ_N is a random measure since X_N is a random matrix. Wigner showed that the limiting behaviour of μ_N does not depend on the details of the random variables which make up X_N. In [Wigner 1958], he made the following hypotheses:

(1) Each $X_N(ij)$ has a symmetric distribution.

(2) Each $X_N(ij)$ has finite moments of all orders, each of which is bounded by a constant independent of N, i, j.

(3) The variance of $X_N(ij)$ is $1/N$.

Wigner proved that, under these hypotheses, μ_N converges weakly to the semicircle law which now bears his name. We will see a proof of Wigner's theorem for random matrices with (complex) Gaussian entries in Lecture Three. The universality of the spectral structure of real and complex Wigner matrices holds at a much finer level, and under much weaker hypotheses, both at the edges of the semicircle [Soshnikov 1999] and in the bulk [Erdős et al. 2011; Tao and Vu 2011].

1.7. ***Noncrossing independence.*** Our quest for the noncrossing Gaussian has brought us into contact with interesting objects (random permutations, elliptic curves, random matrices) and the limit laws which govern them (Tracy–Widom distribution, Sato–Tate distribution, Wigner semicircle distribution). This motivates us to continue developing the rudiments of noncrossing probability theory — perhaps we have hit on a framework within which these objects may be studied.

Our next step is to introduce a notion of noncrossing independence. We know that classical independence is characterized by the vanishing of mixed cumulants. Imitating this, we will define noncrossing independence via the vanishing of mixed noncrossing cumulants. Like classical mixed cumulants, the noncrossing mixed cumulant functionals are defined recursively via the multilinear extension of the noncrossing moment-cumulant formula,

$$m_n(X_1, \dots, X_n) = \sum_{\pi \in \mathsf{NC}(n)} \prod_{B \in \pi} \kappa_{|B|}(X_i : i \in B).$$

The recurrence

$$\kappa_n(X_1, \dots, X_n) = m_n(X_1, \dots, X_n) - \sum_{\pi \in \mathsf{NC}(n)} \prod_{B \in \pi} \kappa_{|B|}(X_i : i \in B)$$

and induction establish that $\kappa_n(X_1, \dots, X_n)$ is a symmetric multilinear function of its arguments. Two random variables X, Y are said to be noncrossing independent if their mixed noncrossing cumulants vanish:

$$\begin{aligned}
&\kappa_2(X, Y) = 0, \\
&\kappa_3(X, X, Y) = \kappa_3(X, Y, Y) = 0, \\
&\kappa_4(X, X, X, Y) = \kappa_4(X, X, Y, Y) = \kappa_4(X, Y, Y, Y) = 0, \\
&\qquad \vdots
\end{aligned}$$

An almost tautological consequence of this definition is that

$$X, Y \text{ noncrossing independent} \implies \kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y) \;\; \forall n \geq 1.$$

Thus, just as classical cumulants linearize the addition of classically independent random variables,

noncrossing cumulants linearize addition
of noncrossing independent random variables.

We can also note that the semicircular random variable X, whose noncrossing cumulant sequence is $0, 1, 0, 0, \dots$, plays the role of the standard Gaussian with respect to this new notion of independence. For example, since noncrossing cumulants linearize noncrossing independence, the sum of two noncrossing independent semicircular random variables is a semicircular random variable of variance two. The noncrossing analogue of the central limit theorem asserts that, if $X_1, X_2, \dots$ is a sequence of noncrossing independent and identically distributed random variables with mean zero and variance one, then the moments of

$$S_N = \frac{X_1 + \cdots + X_N}{\sqrt{N}}$$

converge to the moments of the standard semicircular X as $N \to \infty$. The proof of this fact is identical to the proof of the classical central limit theorem given above, except that classical cumulants are replaced by noncrossing cumulants.

Of course, we don't really know what noncrossing independence means. For example, if X and Y are noncrossing independent, is it true that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$? The answer is yes, since classical and noncrossing mixed cumulants agree up to and including order three,

$$c_1(X) = \kappa_1(X), \quad c_2(X, Y) = \kappa_2(X, Y), \quad c_3(X, Y, Z) = \kappa_3(X, Y, Z).$$

But what about higher order mixed moments?

We observed above that, in the classical case, vanishing of mixed cumulants allows us to recover the familiar algebraic identities governing the expectation of independent random variables. We do not have a priori knowledge of the algebraic identities governing the expectation of noncrossing independent random variables, so we must discover them using the vanishing of mixed noncrossing cumulants. Let us see what this implies for the mixed moment $m_4(X, X, Y, Y) = \mathbb{E}[X^2Y^2]$. Referring to Figure 11 we see that in this case the noncrossing moment-cumulant formula reduces to

$$\begin{aligned} m_4(X, X, Y, Y) &= \kappa_2(X, X)\kappa_2(Y, Y) + \kappa_2(X, X)\kappa_1(Y)\kappa_1(Y) \\ &\quad + \kappa_2(Y, Y)\kappa_1(X)\kappa_1(X) + \kappa_1(X)\kappa_1(X)\kappa_1(Y)\kappa_1(Y), \end{aligned}$$

which is exactly the formula we obtained for classically independent random variables using the classical moment-cumulant formula.

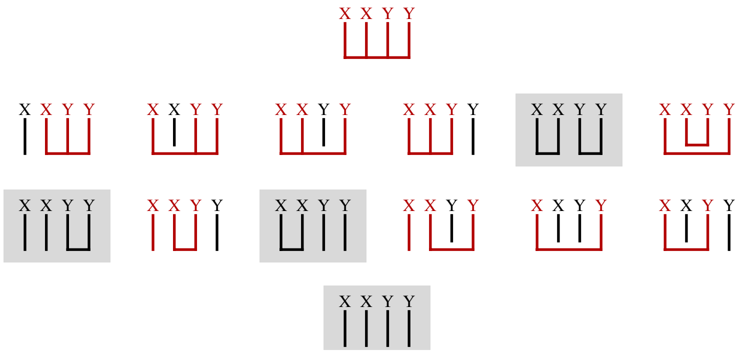

Figure 11. Graphical evaluation of $m_4(X, X, Y, Y)$ using noncrossing cumulants.

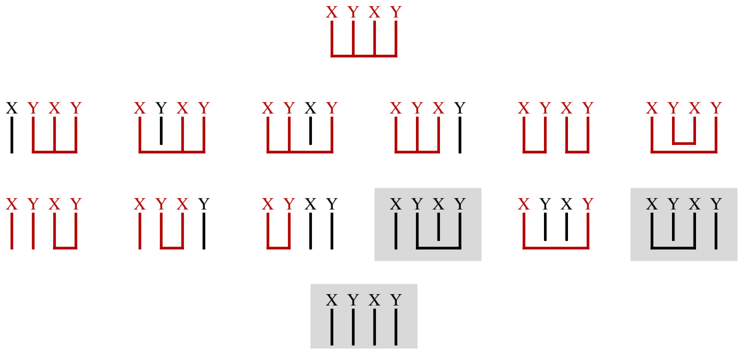

Figure 12. Graphical evaluation of $m_4(X, Y, X, Y)$ using noncrossing cumulants.

However, when we use the noncrossing moment-cumulant formula to evaluate the same mixed moment with its arguments permuted, we instead get

$$m_4(X, Y, X, Y) = \kappa_2(X, X)\kappa_1(Y)\kappa_1(Y) \\ + \kappa_2(Y, Y)\kappa_1(X)\kappa_1(X) + \kappa_1(X)\kappa_1(X)\kappa_1(Y)\kappa_1(Y);$$

see Figure 12. Since $m_4(X, X, Y, Y) = m_4(X, Y, X, Y)$, we are forced to conclude that the two expressions obtained are equal, which in turn forces

$$\kappa_2(X, X)\kappa_2(Y, Y) = 0.$$

Thus, if X, Y are noncrossing independent random variables, at least one of them must have vanishing variance, and consequently must be almost surely constant. The converse is also true — one can show that a (classical or noncrossing) mixed cumulant vanishes if any of its entries are constant random variables. So we have classified pairs of noncrossing independent random variables: they look like $\{X, Y\} = \{\text{arbitrary, constant}\}$. Such pairs of random variables are of no interest from a probabilistic perspective. It would seem that noncrossing probability is a dead end.

1.8. ***The medium is the message.*** If Ω is a compact Hausdorff space then the algebra $\mathcal{A}(\Omega)$ of continuous functions $X : \Omega \to \mathbb{C}$ is a commutative C^*-algebra. This means that in addition to its standard algebraic structure (pointwise addition, multiplication and scalar multiplication of functions) $\mathcal{A}(\Omega)$ is equipped with a norm satisfying the Banach algebra axioms and an antilinear involution which is compatible with the norm, $\|X^*X\| = \|X\|^2$. The norm comes from the topology of the source, $\|X\| = \sup_\omega |X(\omega)|$, and the involution comes from the conjugation automorphism of the target, $X^*(\omega) = \overline{X(\omega)}$. Conversely, a famous theorem of Israel Gelfand asserts that any unital commutative C^*-algebra $\mathcal{A}$ can be realized as the algebra of continuous functions on a compact Hausdorff space $\Omega(\mathscr{A})$ in an essentially unique way. In fact, $\Omega(\mathscr{A})$ may be constructed as the set of maximal ideals of $\mathscr{A}$ equipped with a suitable topology. The associations $\Omega \mapsto \mathscr{A}(\Omega)$ and $\mathscr{A} \mapsto \Omega(\mathscr{A})$ are contravariantly functorial and set up a dual equivalence between the category of compact Hausdorff spaces and the category of unital commutative C^*-algebras.

There are many situations in which one encounters a category of spaces dually equivalent to a category of algebras. In a wonderful book [Nestruev 2003], the mathematicians collectively known as Jet Nestruev develop the theory of smooth real manifolds entirely upside-down: the theory is built in the dual algebraic category, whose objects Nestruev terms smooth complete geometric $\mathbb{R}$-algebras, and then exported to the geometric one by a contravariant functor. In many situations, given a category of spaces dually equivalent to a category of algebras it pays to shift our stance and view the algebraic category as primary. In particular, the algebraic point of view is typically easier to generalize. This is the paradigm shift driving Alain Connes' noncommutative geometry programme, and the reader is referred to [Connes 1994] for much more information.

This paradigm shift is precisely what is needed in order to salvage noncrossing probability theory. In probability theory, the notion of space is that of a Kolmogorov triple $(\Omega, \mathscr{F}, P)$ which models the probability to observe a stochastic system in a given state or collection of states. The dual algebraic object associated to a Kolmogorov triple is $L^\infty(\Omega, \mathscr{F}, P)$, the algebra of essentially bounded complex random variables $X : \Omega \to \mathbb{C}$. Just like in the case of continuous functions on a compact Hausdorff space, this algebra has a very special structure: it is a commutative von Neumann algebra equipped with a unital faithful tracial state, $\tau[X] = \int_\Omega X \, \mathrm{d}P$. Moreover, there is an analogue of Gelfand's theorem in this setting which says that any commutative von Neumann algebra can be realized as the algebra of bounded complex random variables on a Kolmogorov triple in an essentially unique way. This is the statement that the categories of Kolmogorov triples and commutative von Neumann algebras are dual equivalent.

Noncrossing independence was rendered trivial by the commutativity of random variables. We can rescue it from the abyss by following the lead of noncommutative geometry and dropping commutativity in the dual category: we shift our stance and define a noncommutative probability space to be a pair $(\mathscr{A}, \tau)$ consisting of a possibly noncommutative complex associative unital algebra $\mathscr{A}$ together with a unital linear functional $\tau : \mathscr{A} \to \mathbb{C}$. If we reinstate commutativity and insist that $\mathscr{A}$ is a von Neumann algebra and τ a faithful tracial state, we are looking at essentially bounded random variables on a Kolmogorov triple, but a general noncommutative probability space need not be an avatar of any classical probabilistic entity.

As a nod to the origins of this definition, and in order to foster analogies with classical probability, we refer to the elements of $\mathcal{A}$ as random variables and call τ the expectation functional. This prompts some natural questions. Before this subsection we only discussed real random variables — complex numbers crept in with the abstract nonsense. What is the analogue of the notion of real random variable in a noncommutative probability space? Probabilists characterize random variables in terms of their distributions. Can we assign distributions to random variables living in a noncommutative probability space? Is it possible to give meaning to the phrase "the distribution of a bounded real random variable living in a noncommutative probability space is a compactly supported probability measure on the line"? We will deal with some of these questions at the end of Lecture Two. For now, however, we remain in the purely algebraic framework, where the closest thing to the distribution of a random variable $X \in \mathcal{A}$ is its moment sequence $m_n(X) = \tau[X^n]$. As in [Voiculescu et al. 1992, p. 12]:

> The algebraic context is not used in the pursuit of generality, but rather of transparence.

1.9. ***A brief history of the free world.*** Having cast off the yoke of commutativity, we are free — free to explore noncrossing probability in the new framework provided by the noncommutative probability space concept. Noncrossing probability has become *free probability*, and will henceforth be referred to as such. Accordingly, noncrossing cumulants will now be referred to as free cumulants, and noncrossing independence will be termed free independence.

The reader is likely aware that free probability is a flourishing area of contemporary mathematics. This first lecture has been historical fiction, and is essentially an extended version of [Novak and Śniady 2011]. Free probability was not discovered in the context of graph enumeration problems, or by tampering with the cumulant concept, although in retrospect it might have been. Rather, free probability theory was invented by Dan-Virgil Voiculescu in the 1980s in order to address a famous open problem in the theory of von Neumann algebras, the free group factors isomorphism problem. The problem is to determine when the von Neumann algebra of the free group on a generators is isomorphic to the von Neumann algebra of the free group on b generators. It is generally believed that these are isomorphic von Neumann algebras if and only if $a = b$, but this remains an open problem. Free probability theory (and its name) originated in this operator-algebraic context.

Voiculescu's definition of free independence, which was modelled on the free product of groups, is the following: random variables X, Y in a noncommutative probability space $(\mathcal{A}, \tau)$ are said to be freely independent if

$$\tau[f_1(X)g_1(Y) \dots f_k(X)g_k(Y)] = 0$$

whenever $f_1, g_1, \dots, f_k, g_k$ are polynomials such that

$$\tau[f_1(X)] = \tau[g_1(X)] = \cdots = \tau[f_k(X)] = \tau[g_k(Y)] = 0.$$

This should be compared with the definition of classical independence: random variables X, Y in a noncommutative probability space $(\mathcal{A}, \tau)$ are said to be classically independent if they commute, $XY = YX$, and if

$$\tau[f(X)g(Y)] = 0$$

whenever f and g are polynomials such that $\tau[f(X)] = \tau[g(Y)] = 0$. These two definitions are antithetical: classical independence has commutativity built into it, while free independence becomes trivial if commutativity is imposed. Nevertheless, both notions are accommodated within the noncommutative probability space framework.

The precise statement of equivalence between classical independence and vanishing of mixed cumulants is due to Gian-Carlo Rota [1964]. In the 1990s, knowing both of Voiculescu's new free probability theory and Rota's approach to classical probability theory, Roland Speicher made the beautiful discovery that by excising the lattice of set partitions from Rota's foundations and replacing it with the lattice of noncrossing partitions, much of Voiculescu's theory could be recovered and extended by elementary combinatorial methods. In particular, Speicher showed that free independence is equivalent to the vanishing of mixed free cumulants. The combinatorial approach to free probability is exhaustively applied in [Nica and Speicher 2006], while the original analytic approach of Voiculescu is detailed in [Voiculescu et al. 1992].

2. Lecture two: exploring the free world

Lecture One culminated in the notion of a noncommutative probability space and the realization that this framework supports two types of independence: classical independence and free independence. From here we can proceed in several ways. One option is to prove an abstract result essentially stating that these are the only notions of independence which can occur. This result, due to Speicher, places classical and free independence on equal footing. Another possibility is to present concrete problems of intrinsic interest where free independence naturally appears. We will pursue the second route, and examine problems emerging from the theory of random walks on groups which can be recast as questions about free random variables. In the course of solving these problems we will develop the calculus of free random variables and explore the terrain of the free world.

2.1. ***Random walks on the integers.*** The prototypical example of a random walk on a group is the simple random walk on $\mathbf{Z}$: a walker initially positioned at zero tosses a fair coin at each tick of the clock — if it lands heads he takes a step of $+1$, if it lands tails he takes a step of -1. A random walk is said to be recurrent if it returns to its initial position with probability one, and transient if not. Is the simple random walk on $\mathbf{Z}$ recurrent or transient?

Let $\alpha(n)$ denote the number of walks which return to zero for the first time after n steps, and let $\phi(n) = 2^{-n}\alpha(n)$ denote the corresponding probability that the first return occurs at time n. Note that $\alpha(0) = \phi(0) = 0$, and define

$$F(z) = \sum_{n=0}^{\infty} \phi(n) z^n.$$

Then

$$F(1) = \sum_{n=0}^{\infty} \phi(n) \leq 1$$

is the probability we seek. The radius of convergence of $F(z)$ is at least one, and by Abel's theorem

$$F(1) = \lim_{x \to 1} F(x)$$

as x approaches 1 in the interval $[0, 1)$.

Let $\lambda(n)$ denote the number of length n loops on $\mathbf{Z}$ based at 0, and let $\rho(n) = 2^{-n}\lambda(n)$ be the corresponding probability of return at time n (regardless of whether this is the first return or not). Note that $\lambda(0) = \rho(0) = 1$. We have

$$\lambda(n) = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \binom{n}{n/2} & \text{if } n \text{ is even.} \end{cases}.$$

From Stirling's formula, we see that

$$\rho(2k) \sim \frac{1}{\sqrt{\pi k}}$$

as $k \to \infty$. Thus the radius of convergence of

$$R(z) = \sum_{n=0}^{\infty} \rho(n) z^n$$

is 1.

We can decompose the set of loops of given length according to the number of steps taken to the first return. This produces the equation

$$\lambda(n) = \sum_{k=0}^{n} \alpha(k)\lambda(n-k).$$

Equivalently, since all probabilities are uniform,

$$\rho(n) = \sum_{k=0}^{n} \phi(k)\rho(n-k).$$

Summing on z, this becomes the identity

$$R(z) - 1 = F(z)R(z)$$

in the algebra of holomorphic functions on the open unit disc in $\mathbb{C}$. Since $R(z)$ has nonnegative coefficients, it is nonvanishing for $x \in [0, 1)$ and we can write

$$F(x) = 1 - \frac{1}{R(x)}, \quad 0 \leq x < 1.$$

Thus

$$F(1) = \lim_{x \to 1} F(x) = 1 - \frac{1}{\lim_{x \to 1} R(x)}.$$

If $R(1) < \infty$, then by Abel's theorem $\lim_{x\to 1} R(x) = R(1)$ and we obtain $F(1) < 1$. On the other hand, if $R(1) = \infty$, then $\lim_{x\to 1} R(x) = \infty$ and we get $F(1) = 1$. Thus the simple random walk is transient or recurrent according to the convergence or divergence of the series $\sum \rho(n)$. From the Stirling estimate above we find that this sum diverges, so the simple random walk on $\mathbf{Z}$ is recurrent.

2.2. ***Pólya's theorem.*** In the category of abelian groups, coproducts are direct sums:

$$\coprod_{i\in I} \mathbf{G}_i = \bigoplus_{i\in I} \mathbf{G}_i.$$

George Pólya [1921] proved that the simple random walk on

$$\mathbf{Z}^d = \underbrace{\mathbf{Z} \oplus \cdots \oplus \mathbf{Z}}_{d}$$

is recurrent for $d = 1, 2$ and transient for $d > 2$. This striking result can be deduced solely from an understanding of the simple random walk on $\mathbf{Z}$.

Let us give a proof of Pólya's theorem. Let $\lambda_d(n)$ denote the number of length n loops on $\mathbf{Z}^d$ based at 0^d. Let $\rho_d(n)$ denote the probability of return to 0^d after n steps,

$$\rho_d(n) = \frac{1}{(2d)^n}\lambda_d(n).$$

As above, the simple random walk on $\mathbf{Z}^d$ is recurrent if the sum $\sum \rho_d(n)$ diverges, and transient otherwise. Form the loop generating function

$$L_d(z) = \sum_{n=0}^{\infty} \lambda_d(n) z^n.$$

We aim to prove that

$$L_d\left(\frac{1}{2d}\right)=\sum_{n=0}^{\infty}\rho_d(n)$$

diverges for $d=1,2$ and converges for $d>2$.

While the ordinary loop generating function is hard to analyze directly, the exponential loop generating function

$$E_d(z)=\sum_{n=0}^{\infty}\lambda_d(n)\frac{z^n}{n!}$$

is quite accessible. Indeed, as in the last subsection we have

$$\lambda_1(n)=\begin{cases}0 & \text{if } n \text{ is odd,}\\ \binom{n}{n/2} & \text{if } n \text{ is even,}\end{cases}$$

so that

$$E_1(z)=\sum_{k=0}^{\infty}\frac{z^{2k}}{k!\,k!}=I_0(2z)$$

is precisely the modified Bessel function of order zero. Since a loop on $\mathbf{Z}^d$ is just a shuffle of loops on $\mathbf{Z}$, the product formula for exponential generating functions [Stanley 1999] yields

$$E_d(z)=E_1(z)^d=I_0(2z)^d.$$

What we have is the exponential generating function for the loop counts $\lambda_d(n)$, and what we want is the ordinary generating function of this sequence. The integral transform

$$L_f(z)=\int_0^{\infty}f(tz)e^{-t}\,\mathrm{d}t,$$

which looks like the Laplace transform of f but with the z-parameter in the wrong place, converts exponential generating functions into ordinary generating functions. This can be seen by differentiating under the integral sign and using the fact that the moments of the exponential distribution are the factorials,

$$\int_0^{\infty}t^n e^{-t}\,\mathrm{d}t=n!.$$

This trick is constantly used in quantum field theory in connection with Borel summation of divergent series [Etingof 2003]. In particular, we have

$$L_d(z)=\int_0^{\infty}E_d(tz)e^{-t}\,\mathrm{d}t=\int_0^{\infty}I_0(2tz)^d e^{-t}\,\mathrm{d}t.$$

Thus it remains only to show that the integral

$$L_d\left(\frac{1}{2d}\right) = \int_0^\infty I_0\left(\frac{t}{d}\right)^d e^{-t}\,\mathrm{d}t$$

is divergent for $d = 1, 2$ and convergent for $d > 2$. This in turn amounts to understanding the asymptotics of $I_0(t/d)$ as $t \to \infty$ along the real line.

We already encountered Bessel functions in Lecture One, and we know that

$$I_0(t/d) = \frac{1}{\pi}\int_0^\pi e^{t(\frac{\cos\theta}{d})}\,\mathrm{d}\theta.$$

This is an integral of Laplace type,

$$\int_a^b e^{tf(\theta)}\,\mathrm{d}\theta,$$

and Laplace integrals localize as $t \to \infty$ with asymptotics given by the classical steepest descent formula (maximum at an endpoint case),

$$\int_a^b e^{tf(\theta)}\,\mathrm{d}\theta \sim \sqrt{\frac{\pi}{2t|f''(a)|}}e^{tf(a)}.$$

For our integral, this specializes to

$$I_0(t/d) \sim \sqrt{\frac{1}{2\pi^{3/2}t}}e^{t/d}, \quad t \to \infty,$$

from which it follows that $L_d((2d)^{-1})$ diverges or converges according to the divergence or convergence of the integral

$$\int_1^\infty t^{-d/2}\,\mathrm{d}t.$$

This integral diverges for $d = 1, 2$ and converges for $d \geq 3$, which proves Pólya's result. In fact, the probability that the simple random walk on $\mathbf{Z}^3$ returns to its initial position is already less than thirty five percent.

2.3. ***Kesten's problem.*** The category of abelian groups is a full subcategory of the category of groups. In the category of groups, coproduct is free product:

$$\coprod_{i\in I}\mathbf{G}_i = *_{i\in I}\mathbf{G}_i.$$

Thus one could equally well ask about the recurrence or transience of the simple random walk on

$$\mathbf{F}_d = \mathbf{Z} * \cdots * \mathbf{Z},$$

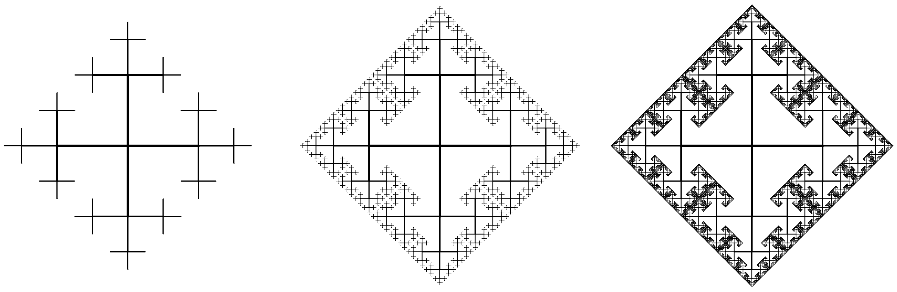

Figure 13. Balls of increasing radius in $\mathbf{F}_2$.

the free group on d generators. Whereas the Cayley graph of the abelian group $\mathbf{Z}^d$ is the $(2d)$-regular hypercubic lattice, the Cayley graph of the free group $\mathbf{F}_d$ is the $(2d)$-regular tree; see Figure 13. What is the free analogue of Pólya's theorem? We will see that the random walk on $\mathbf{F}_d$ can be understood entirely in terms of the random walk on $\mathbf{F}_1 = \mathbf{Z}$, just as in the abelian category. However, the tools we will use are quite different, and the concept of free random variables plays the central role.

The study of random walks on groups was initiated by Harry Kesten in his 1958 Ph.D. thesis, with published results appearing in [Kesten 1959]. A good source of information on this topic, with many pointers to the literature, is Laurent Saloff-Coste's survey article [2001]. Kesten related the behaviour of the simple random walk on a finitely-generated group $\mathbf{G}$ to other properties of $\mathbf{G}$, such as amenability. A countable group is said to be amenable if it admits a finitely additive $\mathbf{G}$-invariant probability measure. The notion of amenability was introduced by John von Neumann in 1929. Finite groups are amenable since they can be equipped with the uniform measure $P(g) = |\mathbf{G}|^{-1}$. For infinite groups the situation is not so clear, and many different characterizations of amenability have been derived. For example, Alain Connes showed that a group is amenable if and only if its von Neumann algebra is hyperfinite. Kesten proved that $\mathbf{G}$ is nonamenable if and only if the probability $\rho_{\mathbf{G}}(n)$ that the simple random walk on $\mathbf{G}$ returns to its starting point at time n decays exponentially in n. We saw above that for $\mathbf{G} = \mathbf{Z}$ the return probability has square root decay, so $\mathbf{Z}$ is amenable. In fact, amenability is preserved by direct sum so all abelian groups are amenable. Is the free group $\mathbf{F}_d$ amenable? Let $\lambda_d(n)$ denote the number of length n loops on $\mathbf{F}_d$ based at id. We will refer to the problem of finding an explicit expression for the loop generating function

$$L_d(z) = 1 + \sum_{n=1}^{\infty} \lambda_d(n) z^n$$

as Kesten's problem. Presumably, if we can obtain an explicit expression for this function then we can read off the asymptotics of $\rho_d(n)$, which is the coefficient of z^n in $L_d(z/2d)$, via the usual methods of singularity analysis of generating functions.

We begin at the beginning: $d=2$. Let A and B denote the generators of $\mathbf{F}_2$, and let $\mathcal{A}=\mathcal{A}[\mathbf{F}_2]$ be the group algebra consisting of formal $\mathbb{C}$-linear combinations of words in these generators and their inverses, A^{-1} and B^{-1}. The identity element of $\mathcal{A}$ is the empty word, which is identified with id $\in \mathbf{F}_2$. Introduce the expectation functional

$$\tau[X] = \text{coefficient of id in } X$$

for each $X \in \mathcal{A}$. Then $(\mathcal{A}, \tau)$ is a noncommutative probability space. A loop id $\to$ id in $\mathbf{F}_2$ is simply a word in A, A^{-1}, B, B^{-1} which reduces to id. Thus the number of length n loops in $\mathbf{F}_2$ is

$$\lambda_2(n) = m_n(X+Y) = \tau[(X+Y)^n],$$

where $X, Y \in \mathcal{A}$ are the random variables

$$X = A + A^{-1}, \quad Y = B + B^{-1}.$$

We see that the loop generating function for $\mathbf{F_2}$ is precisely the moment generating function for the random variable $X+Y$ in the noncommutative probability space $(\mathcal{A}, \tau)$,

$$L_2(z) = 1 + \sum_{n=1}^{\infty} m_n(X+Y) z^n.$$

We want to compute the moments of the sum $X+Y$ of two noncommutative random variables, and what we know are the moments of its summands:

$$m_n(X) = m_n(Y) = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \binom{n}{n/2} & \text{if } n \text{ is even} \end{cases}.$$

Now we make the key observation: the random variables X, Y are freely independent. Indeed, suppose that $f_1, g_1, \ldots, f_k, g_k$ are polynomials such that

$$\tau[f_1(X)] = \tau[g_1(Y)] = \cdots = \tau[f_k(X)] = \tau[g_k(Y)] = 0.$$

This means that $f_i(X) = f_i(A + A^{-1})$ is a Laurent polynomial in A with zero constant term, and $g_j(Y) = g_j(B + B^{-1})$ is a Laurent polynomial in B with zero constant term. Since there are no relations between A and B, an alternating product of polynomials of this form cannot produce any occurrences of the empty word, and we have

$$\tau[f_1(X)g_1(Y)\ldots f_k(X)g_k(Y)] = 0.$$

This is precisely Voiculescu's definition of free independence.

We conclude that the problem of computing $\lambda_2(n)$ is a particular case of the problem of computing the moments $m_n(X+Y)$ of the sum of two free random variables given their individual moments, $m_n(X)$ and $m_n(Y)$. This motivates us to solve a fundamental problem in free probability theory:

> *Given a pair of free random variables X and Y, compute the moments of $X+Y$ in terms of the moments of X and the moments of Y.*

We can, in principle, solve this problem using the fact that free cumulants linearize the addition of free random variables, $\kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y)$. This solution is implemented as the following recursive algorithm.

Input: $\kappa_1(X), \dots, \kappa_{n-1}(X), \kappa_1(Y), \dots, \kappa_{n-1}(Y)$.

***Step 1*:** Compute $m_n(X), m_n(Y)$.

***Step 2*:** Compute $\kappa_n(X), \kappa_n(Y)$ using

$$\kappa_n(X) = m_n(X) - \sum_{\substack{\pi \in \mathsf{NC}(n)\\ b(\pi)\geq 2}} \prod_{B\in\pi} \kappa_{|\beta|}(X)$$

$$\kappa_n(Y) = m_n(Y) - \sum_{\substack{\pi \in \mathsf{NC}(n)\\ b(\pi)\geq 2}} \prod_{B\in\pi} \kappa_{|\beta|}(Y).$$

***Step 3*:** Add:

$$\kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y).$$

***Step 4*:** Compute $m_n(X+Y)$ using

$$m_n(X+Y) = \kappa_n(X+Y) + \sum_{\substack{\pi \in \mathsf{NC}(n)\\ b(\pi\geq 2}} \prod_{B\in\pi} \kappa_{|B|}(X+Y).$$

Output: $m_n(X+Y)$.

This recursive algorithm is conceptually simple but virtually useless as is. In particular, it is not clear how to coax it into computing the loop generating function $L_2(z)$. We need to develop an additive calculus of free random variables which parallels the additive calculus of classically independent random variables.

2.4. ***The classical algorithm.*** If X, Y are classically independent random variables, we can compute the moments of their sum $X+Y$ using the recursive algorithm above, replacing free cumulants with classical cumulants. But this is not what probabilists do in their daily lives. They have a much better algorithm which uses analytic function theory to efficiently handle the recursive nature of the naive algorithm. The classical algorithm associates to X and Y analytic functions $M_X(z)$ and $M_Y(z)$ which have the property that $M_{X+Y}(z) := M_X(z)M_Y(z)$ encodes the moments of $X+Y$ as its derivatives at $z=0$. We will give a somewhat roundabout derivation of this algorithm, which is presented in this way specifically to highlight the analogy with Voiculescu's algorithm presented in the next section.

The classical algorithm for summing two random variables is developed in two stages. In the first stage, the relation between the moments and classical cumulants of a random variable is packaged as an identity in the ring of formal power series $\mathbb{C}[\![z]\!]$. Suppose that $(m_n)_{n=1}^{\infty}$ and $(c_n)_{n=1}^{\infty}$ are two numerical sequences related by the chain of identities

$$m_n = \sum_{\pi \in \mathsf{P}(n)} \prod_{B \in \pi} c_{|B|}, \quad n \geq 1.$$

The π-th term of the sum on the right only depends on the "spectrum" of π, i.e., the integer vector $\Lambda(\pi) = (1^{b_1(\pi)}, 2^{b_2(\pi)}, \dots, n^{b_n(\pi)})$, where $b_i(\pi)$ is the number of blocks of size i in π. We may view $\Lambda(\pi)$ as the Young diagram with b_i rows of length i. Consequently, we can perform a change of variables to push the summation forward onto a sum over Young diagrams with n boxes provided we can compute the "Jacobian" of the map $\Lambda : \mathsf{P}(n) \to \mathsf{Y}(n)$ sending π on its spectrum:

$$m_n = \sum_{b_1+2b_2+\dots+nb_n=n} c_1^{b_1} c_2^{b_2} \dots c_n^{b_n} |\Lambda^{-1}(1^{b_1}, 2^{b_2}, \dots, n^{b_n})|.$$

The volume of the fibre of Λ over any given Young diagram can be explicitly computed to be

$$|\Lambda^{-1}(1^{b_1}, 2^{b_2}, \dots, n^{b_n})| = \frac{n!}{(1!)^{b_1}(2!)^{b_2} \dots (n!)^{b_n} b_1!\, b_2! \dots b_n!},$$

so that we have the chain of identities

$$\frac{m_n}{n!} = \sum_{b_1+2b_2+\dots+nb_n=n} \frac{(c_1/1!)^{b_1}(c_2/2!)^{b_2} \dots (c_n/n!)^{b_n}}{b_1!\, b_2! \dots b_n!}, \quad n \geq 1.$$

We can bundle these identities as a single relation between power series. Summing on z we obtain

$$\begin{aligned} 1 + \sum_{n=1}^{\infty} m_n \frac{z^n}{n!} &= 1 + \sum_{n=1}^{\infty} \Bigg(\sum_{b_1+2b_2+\dots+nb_n=n} \frac{(c_1/1!)^{b_1}(c_2/2!)^{b_2} \dots (c_n/n!)^{b_n}}{b_1!\, b_2! \dots b_n!} \Bigg) z^n \\ &= 1 + \frac{1}{1!}\Bigg(\sum_{n=1}^{\infty} c_n \frac{z^n}{n!}\Bigg)^1 + \frac{1}{2!}\Bigg(\sum_{n=1}^{\infty} c_n \frac{z^n}{n!}\Bigg)^2 + \dots = \exp\Bigg(\sum_{n=1}^{\infty} c_n \frac{z^n}{n!}\Bigg). \end{aligned}$$

We conclude that the chain of moment-cumulant formulas is equivalent to the single identity $M(z) = e^{C(z)}$ in $\mathbb{C}[\![z]\!]$, where

$$M(z) = 1 + \sum_{n=1}^{\infty} m_n \frac{z^n}{n!}, \quad C(z) = \sum_{n=1}^{\infty} c_n \frac{z^n}{n!}$$

This fact is known in enumerative combinatorics as the exponential formula. In other branches of science it goes by other names, such as the polymer expansion formula or the linked cluster theorem. In the physics literature, the exponential formula is often invoked using colourful phrases such as "connected vacuum bubbles exponentiate" [Samuel 1980]. The exponential formula seems to have been first written down precisely by Adolf Hurwitz [1891].

The exponential formula becomes particularly powerful when combined with complex analysis. Suppose that X, Y are classically independent random variables living in a noncommutative probability space $(\mathcal{A}, \tau)$. Suppose moreover that an oracle has given us probability measures μ_X, μ_Y on the real line which behave like distributions for X, Y insofar as

$$\tau[X^n] = \int_{\mathbb{R}} t^n \mu_X(\mathrm{d}t), \quad \tau[Y^n] = \int_{\mathbb{R}} t^n \mu_Y(\mathrm{d}t), \quad n \geq 1.$$

Let us ask for even more, and insist that μ_X, μ_Y are compactly supported. Then the functions[2]

$$M_X(z) = \int_{\mathbb{R}} e^{tz} \mu_X(\mathrm{d}t), \quad M_Y(z) = \int_{\mathbb{R}} e^{tz} \mu_Y(\mathrm{d}t)$$

are entire, and their derivatives can be computed by differentiation under the integral sign. Consequently, we have the globally convergent power series expansions

$$M_X(z) = 1 + \sum_{n=1}^{\infty} m_n(X) \frac{z^n}{n!},$$
$$M_Y(z) = 1 + \sum_{n=1}^{\infty} m_n(Y) \frac{z^n}{n!}.$$

Since $M_X(0) = M_Y(0) = 1$ and the zeros of holomorphic functions are discrete, we can restrict to a complex domain D containing the origin on which $M_X(z)$, $M_Y(z)$ are nonvanishing. Let $\mathrm{Hol}(\mathsf{D})$ denote the algebra of holomorphic functions on D. The following algorithm produces a function $M_{X+Y}(z) \in \mathrm{Hol}(\mathsf{D})$ whose derivatives at $z = 0$ are the moments of $X + Y$.

[2] The restriction of M_X to the real axis, $M_X(-x)$, is the two-sided Laplace transform, while the restriction of M_X to the imaginary axis, $M_X(-iy)$, is the Fourier transform.

Input: μ_X and μ_Y.

***Step 1*:** Compute

$$M_X(z) = \int_{\mathbb{R}} e^{tz}\mu_X(\mathrm{d}t), \;\; M_Y(z) = \int_{\mathbb{R}} e^{tz}\mu_Y(\mathrm{d}t).$$

***Step 2*:** Solve

$$M_X(z) = e^{C_X(z)}, \quad M_Y(z) = e^{C_Y(z)}$$

in $\mathrm{Hol}(\mathsf{D})$ subject to $C_X(0) = C_Y(0) = 0$.

***Step 3*:** Add:

$$C_{X+Y}(z) := C_X(z) + C_Y(z).$$

***Step 4*:** Exponentiate:

$$M_{X+Y}(z) := e^{C_{X+Y}(z)}.$$

Output: $M_{X+Y}(z)$.

In Step 1, we try to compute the integral transforms $M_X(z)$, $M_Y(z)$ in terms of elementary functions, like e^z, $\log(z)$, $\sin(z)$, $\cos(z)$, $\sinh(z)$, $\cosh(z), \ldots$ etc, or other classical functions like Bessel functions, Whittaker functions, or anything else that can be looked up in [Andrews et al. 1999]. This is often feasible if the distributions μ_X, μ_Y have known densities, and we saw some examples in Lecture One.

The equations in Step 2 have unique solutions. The required functions $C_X(z), C_Y(z) \in \mathrm{Hol}(\mathsf{D})$ are the principal branches of the logarithms of $M_X(z)$ and $M_Y(z)$ on D, and can be represented as contour integrals:

$$C_X(z) = \log M_X(z) = \oint_0^z \frac{M_X'(\zeta)}{M_X(\zeta)}\,\mathrm{d}\zeta, \quad C_Y(z) = \log M_Y(z) = \oint_0^z \frac{M_Y'(\zeta)}{M_Y(\zeta)}\,\mathrm{d}\zeta$$

for $z \in \mathsf{D}$. Since log has the usual formal properties associated with the logarithm, if Step 1 outputs a reasonably explicit expression then so will Step 2.

Step 2 is the crux of the algorithm. It is performed precisely to change gears from a moment computation to a cumulant computation. Appealing to the exponential formula, we conclude that the holomorphic functions $C_X(z)$, $C_Y(z)$ passed to Step 3 by Step 2 have Maclaurin series

$$C_X(z) = \sum_{n=1}^{\infty} c_n(X)\frac{z^n}{n!}, \quad C_Y(z) = \sum_{n=1}^{\infty} c_n(Y)\frac{z^n}{n!},$$

where $c_n(X)$, $c_n(Y)$ are the cumulants of X and Y. Since cumulants linearize the addition of independent random variables, the new function $C_{X+Y}(z) := C_X(z)+C_Y(z)$ defined in Step 3 encodes the cumulants of $X+Y$ as its derivatives at $z=0$.

In Step 4 we define a new function $M_{X+Y}(z) \in \mathrm{Hol}(\mathsf{D})$ by $M_{X+Y}(z) := e^{C_{X+Y}(z)}$. The exponential formula and the moment-cumulant formula now combine in the reverse direction to tell us that the Maclaurin series of $M_{X+Y}(z)$ is

$$M_{X+Y}(z) = 1 + \sum_{n=1}^{\infty} m_n(X+Y)\frac{z^n}{n!}.$$

In summary, assuming that X, Y are classically independent random variables living in a noncommutative probability space $(\mathscr{A}, \tau)$ with affiliated distributions μ_X, μ_Y having nice properties, the classical algorithm takes these distributions as input and outputs a function $M_{X+Y}(z)$ analytic at $z=0$ whose derivatives are the moments of $X+Y$. It works by combining the exponential formula and the moment-cumulant formula to convert the moment problem into the (linear) cumulant problem, adding, and then converting back to moments. An optional Step 5 is to extract the distribution μ_{X+Y} from $M_{X+Y}(z)$ using the Fourier inversion formula:

$$\mu_{X+Y}([a,b]) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iat}-e^{-ibt}}{it} M_{X+Y}(it)\,\mathrm{d}t.$$

2.5. ***Voiculescu's algorithm.*** We wish to develop a free analogue of the classical algorithm. Suppose that X, Y are freely independent random variables living in a noncommutative probability space $(\mathscr{A}, \tau)$ possessing compactly supported real distributions μ_X, μ_Y. The free algorithm should take these distributions as input, build a pair of analytic functions which encode the moments of X and Y respectively, and then convolve these somehow to produce a new analytic function which encodes the moments of $X+Y$. A basic hurdle to be overcome is that, even assuming we know how to construct μ_X and μ_Y, we don't know what to do with them. We could repeat Step 1 of the classical algorithm to obtain analytic functions $M_X(z)$, $M_Y(z)$ whose derivatives at $z=0$ are the moments of X and Y. If we then perform Step 2 we obtain analytic functions $C_X(z)$, $C_Y(z)$ whose derivatives encode the classical cumulants of X and Y. But classical cumulants do not linearize addition of free random variables.

The classical algorithm is predicated on the existence of a formal power series identity equivalent to the chain of classical moment-cumulant identities. We need a free analogue of this, namely a power series identity equivalent to the chain of numerical identities

$$m_n = \sum_{\pi \in \mathsf{NC}(n)} \prod_{B\in\pi} \kappa_{|B|}, \quad n \geq 1.$$

Proceeding as in the classical case, rewrite this in the form

$$m_n = \sum_{b_1+2b_2+\dots+nb_n=n} \kappa_1^{b_1}\kappa_2^{b_2}\dots\kappa_n^{b_n}|\Lambda^{-1}(1^{b_1}, 2^{b_2}, \dots, n^{b_n}) \cap \mathsf{NC}(n)|,$$

where as above $\Lambda : \mathsf{P}(n) \to \mathsf{Y}(n)$ is the surjection which sends a partition π with b_i blocks of size i to the Young diagram with b_i rows of length i. Now we have to compute the volume of the fibres of Λ intersected with the noncrossing partition lattice. The solution to this enumeration problem is again known in explicit form,

$$|\Lambda^{-1}(1^{m_1}, 2^{m_2}, \dots, n^{m_n}) \cap \mathsf{NC}(n)| = \frac{n!}{(n+1-(b_1+b_2+\dots+b_n))!b_1!b_2!\dots b_n!}.$$

This formula allows us to obtain the desired power series identity, though the manipulations required are quite involved and require either the use of Lagrange inversion or an understanding of the poset structure of $\mathsf{NC}(n)$. In any event, what ultimately comes out of the computation is the fact that two numerical sequences satisfy the chain of free moment-cumulant identities if and only if the ordinary (not exponential) generating functions

$$L(z) = 1 + \sum_{n=1}^{\infty} m_n z^n, \quad K(z) = 1 + \sum_{n=1}^{\infty} \kappa_n z^n$$

solve the equation

$$L(z) = K(zL(z))$$

in the formal power series ring $\mathbb{C}[\![z]\!]$. This is the free analogue of the exponential formula.

As in the classical case, we wish to turn this formal power series encoding into an analytic encoding. Suppose that X, Y admit distributions μ_X, μ_Y supported in the real interval $[-r, r]$. We then have $|m_n(X)|, |m_n(Y)| \leq r^n$, so the moment generating functions

$$L_X(z) = 1 + \sum_{n=1}^{\infty} m_n(X) z^n, \quad L_Y(z) = 1 + \sum_{n=1}^{\infty} m_n(Y) z^n,$$

are absolutely convergent in the open disc $\mathsf{D}(0, \frac{1}{r})$. One can use the relation between moments and free cumulants to show that the free cumulant generating functions

$$K_X(z) = 1 + \sum_{n=1}^{\infty} \kappa_n(X) z^n, \quad K_Y(z) = 1 + \sum_{n=1}^{\infty} \kappa_n(Y) z^n$$

are also absolutely convergent on a (possibly smaller) neighbourhood of $z = 0$.

However, it turns out that the correct environment for the free algorithm is a neighbourhood of infinity rather than a neighbourhood of zero. This is because what we really want is an integral transform which realizes ordinary generating functions in the same way as the Fourier (or Laplace) transform realizes exponential generating functions. Access to such a transform will allow us to obtain closed forms for generating functions by evaluating integrals, just like in classical probability. Such an object is well-known in analysis, where it goes by the name of the Cauchy (or Stieltjes) transform. The Cauchy transform of a random variable X with real distribution μ_X is

$$G_X(z) = \int_{\mathbb{R}} \frac{1}{z-t} \mu_X(\mathrm{d}t).$$

The Cauchy transform is well-defined on the complement of the support of μ_X, and differentiating under the integral sign shows that $G_X(z)$ is holomorphic on its domain of definition. In particular, if μ_X is supported in $[-r, r]$ then $G_X(z)$ admits the convergent Laurent expansion

$$G_X(z) = \frac{1}{z} \sum_{n=0}^{\infty} \frac{\int t^n \mu_X(\mathrm{d}t)}{z^n} = \sum_{n=0}^{\infty} \frac{m_n(X)}{z^{n+1}}$$

on $|z| > r$. This is an ordinary generating function for the moments of X with z^{-1} playing the role of the formal variable.

To create an interface between the free moment-cumulant formula and the Cauchy transform, we must rewrite the formal power series identity $L(z) = K(zL(z))$ as an identity in $\mathbb{C}((z)) = \operatorname{Quot} \mathbb{C}[\![z]\!]$, the field of formal Laurent series. Introduce the formal Laurent series

$$G(z) = \frac{1}{z} L\left(\frac{1}{z}\right) = \sum_{n=0}^{\infty} \frac{m_n}{z^{n+1}}.$$

The automorphism $z \mapsto \dfrac{1}{z}$ transforms the noncrossing exponential formula into the identity

$$\frac{K(G(z))}{G(z)} = z.$$

Setting

$$V(z) = \frac{K(z)}{z} = \frac{1}{z} + \sum_{n=0}^{\infty} \kappa_{n+1} z^n,$$

this becomes the identity

$$V(G(z)) = z$$

in $\mathbb{C}((z))$.

We have now associated two analytic functions to X. The first is the Cauchy transform $G_X(z)$, which is defined as an integral transform and admits a convergent Laurent expansion in a neighbourhood of infinity in the z-plane. The second is the Voiculescu transform $V_X(w)$, which is defined by the convergent Laurent series

$$V_X(w) = \frac{1}{w} + \sum_{n=0}^{\infty} \kappa_{n+1} w^n$$

in a neighbourhood of zero in the w-plane. The Voiculescu transform is a meromorphic function with a simple pole of residue one at $w=0$. The Voiculescu transform less its principal part, $R_X(w) = V_X(w) - \frac{1}{w}$, is an analytic function known as the R-transform of X. From the formal identities $V(G(z)) = z,\ G(V(w)) = w$ and the asymptotics $G_X(z) \sim \frac{1}{z}$ as $|z| \to \infty$ and $V_X(w) \sim \frac{1}{w}$ as $|w| \to 0$, we expect to find a neighbourhood D_∞ of infinity in the z-plane and a neighbourhood D_0 of zero in the w-plane such that $G_X : \mathsf{D}_\infty \to \mathsf{D}_0$ and $V_X : \mathsf{D}_0 \to \mathsf{D}_\infty$ are mutually inverse holomorphic bijections. The existence of the required domains hinges on identifying regions where the Cauchy and Voiculescu transforms are injective, and this can be established through a complex-analytic argument; see [Mingo and Speicher $\geq$ 2014, Chapter 4].

With these pieces in place, we can state Voiculescu's algorithm for the addition of free random variables.

Input: μ_X and μ_Y.

***Step 1*:** Compute

$$G_X(z) = \int_{\mathbb{R}} \frac{1}{z-t} \mu_X(\mathrm{d}t), \qquad G_Y(z) = \int_{\mathbb{R}} \frac{1}{z-t} \mu_Y(\mathrm{d}t)$$

***Step 2*:** Solve the first Voiculescu functional equations,

$$(G_X \circ V_X)(w) = w, \qquad (G_Y \circ V_Y)(w) = w$$

subject to $V_X(w) \sim \dfrac{1}{w}$ near $w = 0$.

***Step 3*:** Remove principal part:

$$R_X(w) = V_X(w) - \frac{1}{w}, \qquad R_Y(w) = V_Y(w) - \frac{1}{w};$$

add:

$$R_{X+Y}(w) := R_X(w) + R_Y(w);$$

restore principal part:

$$V_{X+Y}(w) := R_{X+Y}(w) + \frac{1}{w}.$$

***Step 4*:** Solve the second Voiculescu functional equation,

$$(V_{X+Y} \circ G_{X+Y})(z) = z,$$

subject to $G_{X+Y}(z) \sim \dfrac{1}{z}$ near $z = \infty$.

Output: $G_{X+Y}(z)$.

Voiculescu's algorithm is directly analogous to the classical algorithm presented in the previous section. The analogy can be succinctly summarized by saying that

$$\boxed{\textit{the } R\textit{-transform is the free analogue of the log of the Fourier transform.}}$$

In Step 1, we try to compute the integral transforms $G_X(z)$, $G_Y(z)$ in terms of elementary functions.

Step 2 changes gears from a moment computation to a cumulant computation. Since free cumulants linearize the addition of free random variables, the new function $V_{X+Y}(w) := R_X(w) + R_Y(w) + \frac{1}{w}$ defined in Step 3 encodes the free cumulants of $\kappa_n(X+Y)$ as its Laurent coefficients of nonnegative degree.

In Step 4 we define a new function $G_{X+Y}(z)$ by solving the second Voiculescu functional equation. The free exponential formula and the free moment-cumulant formula combine in the reverse direction to tell us that the Laurent series of $G_{X+Y}(z)$ is

$$G_{X+Y}(z) = \sum_{n=0}^{\infty} \frac{m_n(X+Y)}{z^{n+1}}.$$

An optional fifth step is to extract the distribution μ_{X+Y} from $G_{X+Y}(z)$ using the Stieltjes inversion formula:

$$\mu_{X+Y}(\mathrm{d}t) = -\frac{1}{\pi} \lim_{\varepsilon \to 0} \Im G_{X+Y}(t + i\varepsilon).$$

2.6. ***Solution of Kesten's problem.*** Our motivation for building up the additive theory of free random variables came from Kesten's problem: explicitly determine the loop generating function of the free group $\mathbf{F}_2$, and more generally of the free group $\mathbf{F}_d,\ d \geq 2$. This amounts to computing the moment generating function

$$L_d(z) = 1 + \sum_{n=1}^{\infty} m_n(S_d) z^d$$

of the sum

$$S_d = X_1 + \cdots + X_d$$

of fid (free identically distributed) random variables with moments

$$\tau[X_i^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \binom{n}{n/2} & \text{if } n \text{ is even.} \end{cases}$$

Voiculescu's algorithm gives us the means to obtain this generating function provided we can feed it the required input, namely a compactly supported probability measure on $\mathbb{R}$ with moment sequence

$$0, \tbinom{2}{1}, 0, \tbinom{4}{2}, 0, \tbinom{6}{3}, 0, \ldots .$$

As we saw above, the exponential generating function of this moment sequence,

$$M(z) = \sum_{k=0}^{\infty} \frac{z^{2k}}{k!\,k!} = I_0(2z),$$

coincides with the modified Bessel function of order zero. From the integral representation

$$I_0(2z) = \frac{1}{\pi} \int_0^{\pi} e^{2(\cos\theta)z}\, \mathrm{d}\theta$$

we conclude that a random variable X with odd moments zero and even moments the central binomial coefficients is given by $X = 2\cos(Y)$, where Y has uniform distribution over $[0, \pi]$. Making the same change of variables that we did in Lecture One, we obtain

$$M_X(z) = \frac{1}{\pi} \int_{-2}^{2} e^{tz} \frac{1}{\sqrt{4-t^2}}\, \mathrm{d}t,$$

so that μ_X is supported on $[-2, 2]$ with density

$$\mu_X(\mathrm{d}t) = \frac{1}{\pi\sqrt{4-t^2}}\, \mathrm{d}t.$$

This measure is known as the arcsine distribution because its cumulative distribution function is

$$\int_{-2}^{x} \mu_X(\mathrm{d}t) = \frac{1}{2} + \frac{\arcsin \frac{x}{2}}{\pi}.$$

So to obtain the loop generating function $L_2(z)$ for the simple random walk on $\mathbf{F}_2$, we should run Voiculescu's algorithm with input $\mu_X = \mu_Y =$ arcsine.

Let us warm up with an easier computation. Suppose that X, Y are not fid arcsine random variables, but rather fid ± 1-Bernoulli random variables:

$$\mu_X = \mu_Y = \tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_{+1}.$$

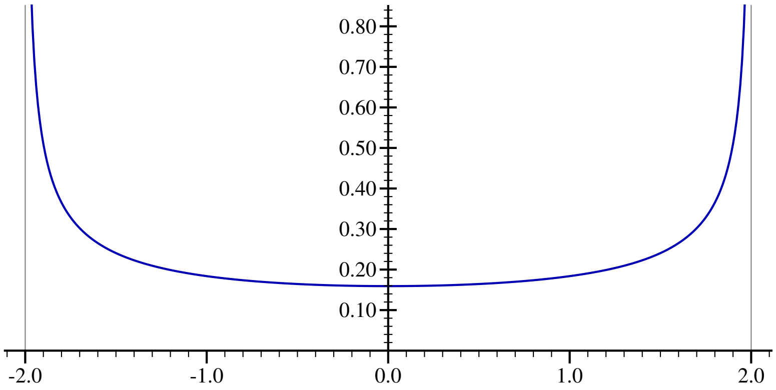

Figure 14. The arcsine density.

We will use Voiculescu's algorithm to obtain the distribution of $X+Y$. If X, Y were classically iid Bernoullis, we would of course obtain the binomial distribution

$$\mu_{X+Y} = \tfrac{1}{4}\delta_{-2} + \tfrac{1}{2}\delta_0 + \tfrac{1}{4}\delta_{+2}$$

giving the distribution of the simple random walk on $\mathbf{Z}$ at time two. The result is quite different in the free case.

Step 1. Obtain the Cauchy transform:

$$G_X(z) = G_Y(z) = \frac{1}{2}\left(\frac{1}{z+1} + \frac{1}{z-1}\right) = \frac{z}{z^2-1} = \sum_{n=0}^{\infty} \frac{1}{z^{2n+1}}.$$

Step 2. Solve the first Voiculescu functional equation. From Step 1, this is

$$wV^2(w) - V(w) - w = 0,$$

which has roots

$$\frac{1+\sqrt{1+4w^2}}{2w} = \frac{1}{w} + w - w^3 + 2w^5 - 5w^7 + \cdots,$$
$$\frac{1-\sqrt{1+4w^2}}{2w} = -w + w^3 - 2w^5 + \cdots.$$

We identify the first of these as the Voiculescu transform $V_X(w) = V_Y(w)$.

Step 3. Compute the R-transform:

$$R_X(w) = R_Y(w) = \frac{1+\sqrt{1+4w^2}}{2w} - \frac{1}{w} = \frac{\sqrt{1+4w^2}-1}{2w},$$

and sum to obtain

$$R_{X+Y}(w) = R_X(w) + R_Y(w) = \frac{\sqrt{1+4w^2}-1}{w}.$$

Now restore the principal part:

$$V_{X+Y}(w) = R_{X+Y}(w) + \frac{1}{w} = \frac{\sqrt{1+4w^2}}{w}.$$

Step 4. Solve the second Voiculescu functional equation. From Step 3, this is the equation

$$\frac{\sqrt{1+4G(z)^2}}{G(z)} = z,$$

which has roots

$$\frac{\pm 1}{\sqrt{z^2-4}} = \frac{\pm 1}{z} + \frac{\pm 2}{z^3} + \frac{\pm 6}{z^5} + \frac{\pm 20}{z^7} + \frac{\pm 70}{z^9} + \frac{\pm 252}{z^{11}} + \cdots .$$

The positive root is identified as $G_{X+Y}(z)$.

Finally, we perform the optional fifth step to recover the distribution μ_{X+Y} whose Cauchy transform is $G_{X+Y}(z)$. This can be done in two ways. First, we could notice that the nonzero Laurent coefficients of G_{X+Y} are the central binomial coefficients $\binom{2k}{k}$, and we just determined that these are the moments of the arcsine distribution. Alternatively we could use Stieltjes inversion:

$$\mu_{X+Y}(\mathrm{d}t) = -\frac{1}{\pi} \lim_{\varepsilon \to 0} \frac{1}{\sqrt{(t+i\varepsilon)^2-4}} = -\frac{1}{\pi} \Im \frac{1}{\sqrt{t^2-4}} = \frac{1}{\pi\sqrt{4-t^2}} \delta_{|t|\leq 2}.$$

We conclude that the sum of two fid Bernoulli random variables has arcsine distribution. Note the surprising feature that the outcome of a free coin toss has continuous distribution over $[-2, 2]$. More generally, we can say that the sum

$$S_d = X_1 + \cdots + X_{2d}$$

of $2d$ fid ± 1-Bernoulli random variables, i.e., the sum of $2d$ free coin tosses, encodes all information about the simple random walk on $\mathbf{F}_d$ in its moments.

Let us move on to the solution of Kesten's problem for $\mathbf{F}_2$. Here X, Y are fid arcsine random variables.

Step 1. The Cauchy transform $G_X(z) = G_Y(z)$ is the output of our last application of the algorithm, namely

$$G_X(z) = G_Y(z) = \frac{1}{\sqrt{z^2-4}}.$$

Step 2. Solve the first Voiculescu functional equation to obtain

$$V_X(w) = V_Y(w) = \frac{\sqrt{1+4w^2}}{w} = \frac{1}{w} + 2w - 2w^3 + \cdots .$$

Step 3. Switch to R-transforms, add, switch back to get the Voiculescu transform of $X+Y$:

$$V_{X+Y}(w) = \frac{2\sqrt{1+4w^2}-1}{z} = \frac{1}{w} + 4w - 4w^3 + \cdots .$$

Step 4. Solve the second Voiculescu functional equation to obtain

$$G_{X+Y}(z) = \frac{-z+2\sqrt{z^2-12}}{z^2-16} = \frac{1}{z} + \frac{4}{z^3} + \frac{28}{z^5} + \frac{232}{z^7} + \frac{2092}{z^9} + \cdots .$$

We can now calculate the loop generating function for $\mathbf{F}_2$:

$$\begin{aligned} L_2(z) &= \frac{1}{z} G_{X+Y}\left(\frac{1}{z}\right) \\ &= \frac{-1+2\sqrt{1-12z^2}}{1-16z^2} = 1 + 4z^2 + 28z^4 + 232z^6 + 2092z^8 + \cdots . \end{aligned}$$

More generally, we can run through the above steps for general d to obtain the loop generating function

$$L_d(z) = \frac{-(d-1)+d\sqrt{1-4(2d-1)z^2}}{1-16z^2}$$

for the free group $\mathbf{F}_d$, $d \geq 2$, which in turn leads to the probability generating function

$$L_d\left(\frac{z}{2d}\right) = \frac{-(d-1)+d\sqrt{1-(2d-1)(\frac{z}{d})^2}}{1-4(\frac{z}{d})^2}.$$

Applying standard methods from analytic combinatorics [Flajolet and Sedgewick 2009], this expression leads to the asymptotics

$$\rho_d(n) \sim \mathrm{const}_d \cdot n^{-3/2} \left(\frac{2\sqrt{d}}{d+1}\right)^n$$

for the return probability of the simple random walk on $\mathbf{F}_d$, $d \geq 2$. From this we can conclude that the simple random walk on $\mathbf{F}_d$ is transient for all $d \geq 2$, and indeed that $\mathbf{F}_d$ is nonamenable for all $d \geq 2$.

2.7. ***Spectral measures and free convolution.*** Voiculescu's algorithm outputs a function $G_{X+Y}(z)$ which encodes the moments of the sum of two freely independent random variables X and Y. As input, it requires a pair of compactly supported real measures μ_X, μ_Y which act as distributions for X and Y in the sense that

$$\tau[X^n] = \int_{\mathbb{R}} t^n \mu_X(\mathrm{d}t), \quad \tau[Y^n] = \int_{\mathbb{R}} t^n \mu_Y(\mathrm{d}t).$$

In our applications of Voiculescu's algorithm we were able to find such measures by inspection. Nevertheless, it is of theoretical and psychological importance to determine sufficient conditions guaranteeing the existence of measures with the required properties.

If $X : \Omega \to \mathbb{C}$ is a random variable defined on a Kolmogorov triple $(\Omega, \mathscr{F}, P)$, its distribution μ_X is the pushforward of P by X,

$$\mu_X(B) = (X_* P)(B) = P(X^{-1}(B))$$

for any Borel (or Lebesgue) set $B \subseteq \mathbb{C}$. One has the general change of variables formula

$$\mathbb{E}[f(X)] = \int_{\mathbb{C}} f(z) \mu_X(\mathrm{d}z)$$

for any reasonable $f : \mathbb{C} \to \mathbb{C}$. If X is essentially bounded and real-valued, μ_X is compactly supported in $\mathbb{R}$. As a random variable X living in an abstract noncommutative probability space $(\mathscr{A}, \tau)$ is not a function, one must obtain μ_X by some other means.

The existence of distributions is too much to expect within the framework of a noncommtative probability space, which is a purely algebraic object. We need to inject some analytic structure into $(\mathscr{A}, \tau)$. This is achieved by upgrading $\mathscr{A}$ to a $*$-algebra, i.e., a complex algebra equipped with a map $* : \mathscr{A} \to \mathscr{A}$ satisfying

$$(X^*)^* = X, \quad (\alpha X + \beta Y)^* = \overline{\alpha} X^* + \overline{\beta} Y^*, \quad (XY)^* = Y^* X^*.$$

This map, which is an abstraction of complex conjugation, is required to be compatible with the expectation τ in the sense that

$$\tau[X^*] = \overline{\tau[X]}.$$

A noncommutative probability space equipped with this extra structure is called a noncommutative $*$-probability space.

In the framework of a $*$-probability space we can single out a class of random variables analogous to real random variables in classical probability. These are the fixed points of $*$, $X^* = X$. A random variable with this property is called self-adjoint. Self-adjoint random variables have real expected values,

$\tau[X] = \tau[X^*] = \overline{\tau[X]}$, and more generally $\tau[f(X)] \in \mathbb{R}$ for any polynomial f with real coefficients.

The identification of bounded random variables requires one more upgrade. Given a $*$-probability space $(\mathscr{A}, \tau)$, we can introduce a Hermitian form $B : \mathscr{A} \times \mathscr{A} \to \mathbb{C}$ defined by

$$B(X, Y) = \tau[XY^*].$$

If we require that τ has the positivity property $\tau[XX^*] \geq 0$ for all $X \in \mathscr{A}$, then we obtain a seminorm

$$\|X\| = B(X, X)^{1/2}$$

on $\mathscr{A}$, and we can access the Cauchy–Schwarz inequality

$$|B(X, Y)| \leq \|X\|\|Y\|.$$

Once we have Cauchy–Schwarz, we can prove the monotonicity inequalities

$$\begin{aligned} |\tau[X]| &\leq |\tau[X^2]|^{1/2} \leq |\tau[X^4]|^{1/4} \\ |\tau[X^3]| &\leq |\tau[X^4]|^{1/4} \leq |\tau[X^6]|^{1/6} \\ |\tau[X^5]| &\leq |\tau[X^6]|^{1/6} \leq |\tau[X^8]|^{1/8} \\ &\vdots \end{aligned}$$

from which the chain of inequalities

$$|\tau[X]| \leq |\tau[X^2]|^{1/2} \leq |\tau[X^4]|^{1/4} \leq |\tau[X^6]|^{1/6} \leq |\tau[X^8]|^{1/8} \leq \cdots$$

can be extracted. From this we conclude that the limit

$$\rho(X) := \lim_{k \to \infty} |\tau[X^{2k}]|^{1/(2k)}$$

exists in $\mathbb{R}_{\geq 0} \cup \{\infty\}$. This limit is called the spectral radius of X. A random variable $X \in \mathscr{A}$ is said to be bounded if its spectral radius is finite, $\rho(X) < \infty$.

In the framework of a noncommutative $*$-probability space $(\mathscr{A}, \tau)$ with nonnegative expectation, bounded self-adjoint random variables play the role of essentially bounded real-valued random variables in classical probability theory. With some work, one may deduce from the Riesz representation theorem that to each bounded self-adjoint X corresponds a unique Borel measure μ_X supported in $[-\rho(X), \rho(X)]$ such that

$$\tau[f(X)] = \int_{\mathbb{R}} f(t)\mu_X(\mathrm{d}t)$$

for all polynomial functions $f : \mathbb{C} \to \mathbb{C}$. The details of this argument, in which a reverse-engineered Cauchy transform plays the key role, are given in Tao's notes [Tao 2010]. The measure μ_X is often called the spectral measure of X, but we will refer to it as the distribution of X. There is also a converse to this result: given any compactly supported measure μ on $\mathbb{R}$, there exists a bounded self-adjoint random variable X living in some noncommutative $*$-probability space $(\mathscr{A}, \tau)$ whose distribution is μ. Consequently, given two compactly supported real probability measures μ, ν we may define a new measure $\mu \boxplus \nu$ as "the distribution of the random variable $X+Y$, where X and Y are freely independent bounded self-adjoint random variables with distributions μ and ν, respectively." Since the sum of two bounded self-adjoint random variables is again bounded self-adjoint, $\mu \boxplus \nu$ is another compactly supported real probability measure. Moreover, $\mu \boxplus \nu$ does not depend on the particular random variables chosen to realize μ and ν. Thus we get a bona fide binary operation $\boxplus$ on the set of compactly supported real measures, which is known as the additive free convolution. For example, we computed above that

$$\text{Bernoulli} \boxplus \text{Bernoulli} = \text{Arcsine}.$$

The additive free convolution of measures is induced by the addition of free random variables. As such, it is the free analogue of the classical convolution of measures induced by the addition of classically independent random variables. Like classical convolution, free convolution can be defined for unbounded measures, but this requires more work [Bercovici and Voiculescu 1993].

2.8. ***Free Poisson limit theorem.*** Select positive real numbers λ and α. Consider the measure

$$\mu_N = (1 - \frac{\lambda}{N})\delta_0 + \frac{\lambda}{N}\delta_\alpha$$

which consists of an atom of mass $1 - \frac{\lambda}{N}$ placed at zero and an atom of mass $\frac{\lambda}{N}$ placed at α. For N sufficiently large, μ_N is a probability measure. Its moment sequence is

$$m_n(\mu_N) = \frac{\lambda}{N}\alpha^n, \quad n \geq 1.$$

The N-fold classical convolution of μ_N with itself,

$$\mu_N^{*N} = \underbrace{\mu_N * \cdots * \mu_N}_{N},$$

converges weakly to the Poisson measure of rate λ and jump size α as $N \to \infty$. This is a classical limit theorem in probability known as the Poisson limit theorem, or the law of rare events.

Let us obtain a free analogue of the Poisson limit theorem. This should be a limit law for the iterated free convolution

$$\mu_N^{\boxplus N} = \underbrace{\mu_N \boxplus \cdots \boxplus \mu_N}_{N}.$$

From the free moment-cumulant formula, we obtain the estimate

$$\kappa_n(\mu_N) = m_n(\mu_N) + O\left(\frac{1}{N^2}\right) = \frac{\lambda}{N}\alpha^n + O\left(\frac{1}{N^2}\right).$$

Since free cumulants linearize free convolution, we have

$$\kappa_n(\mu_N^{\boxplus N}) = N\kappa_n(\mu_N) = \lambda\alpha^n + O\left(\frac{1}{N}\right).$$

Thus

$$\lim_{N\to\infty} \kappa_n(\mu_N) = \lambda\alpha^n,$$

and it remains to determine the measure μ with this free cumulant sequence. The Voiculescu transform of μ is

$$V_\mu(w) = \frac{1}{w} + \sum_{n=0}^{\infty} \lambda\alpha^{n+1}w^n = \frac{1}{w} + \frac{\lambda\alpha}{1-\alpha w},$$

so the second Voiculescu functional equation $V_\mu(G_\mu(z)) = z$ yields

$$\frac{1}{G_\mu(z)} + \frac{\lambda\alpha}{1-\alpha G_\mu(z)} = z.$$

This equation has two solutions, and the one which behaves like $1/z$ for $|z| \to \infty$ is the Cauchy transform of μ. We obtain

$$G_\mu(z) = \frac{z + \alpha(1-\lambda) - \sqrt{(z-\alpha(1+\lambda))^2 - 4\lambda\alpha^2}}{2\alpha z}.$$

Applying Stieltjes inversion, we find that the density of μ is given by

$$\mu(\mathrm{d}t) = \begin{cases} (1-\lambda)\delta_0 + \lambda m(t)\,\mathrm{d}t, & 0 \le \lambda \le 1 \\ m(t)\,\mathrm{d}t, & \lambda > 1 \end{cases}$$

where

$$m(t) = \frac{1}{2\pi\alpha t}\sqrt{4\lambda\alpha^2 - (t-\alpha(1+\lambda))^2}.$$

This measure is known as the Marchenko–Pastur distribution after the Ukrainian mathematical physicists Vladimir Marchenko and Leonid Pastur, who discovered it in their study of the asymptotic eigenvalue distribution of a certain class of random matrices.

2.9. ***Semicircle flow.*** Given $r > 0$, let μ_r be the semicircular measure of radius r:

$$\mu_r(\mathrm{d}t) = \frac{2}{\pi r^2}\sqrt{r^2 - t^2}\,\mathrm{d}t.$$

Taking $r = 2$ yields the standard semicircular distribution. Let μ be an arbitrary compactly supported probability measure on $\mathbb{R}$. The function

$$f_\mu : \{\text{positive real numbers}\} \to \{\text{compactly supported real measures}\}$$

defined by

$$f_\mu(r) = \mu \boxplus \mu_r$$

is called the semicircle flow. The semicircle flow has very interesting dynamics: in one of his earliest articles on free random variables, Voiculescu [1986] showed that

$$\frac{\partial G(r,z)}{\partial r} + G(r,z)\frac{\partial G(r,z)}{\partial z} = 0,$$

where $G(r,z)$ is the Cauchy transform of $f_\mu(r) = \mu \boxplus \mu_r$. Thus the free analogue of the heat equation is the complex inviscid Burgers equation. For a detailed analysis of the semicircle flow, see [Biane 1997].

3. Lecture three: modelling the free world

Free random variables are of interest for many reasons. First and foremost, Voiculescu's free probability theory is an intrinsically appealing subject worthy of study from a purely esthetic point of view. Adding to this are the many remarkable connections between free probability and other parts of mathematics, including operator algebras, representation theory, and random matrix theory. This lecture is an exposition of Voiculescu's discovery that random matrices provide asymptotic models of free random variables. We follow the treatment of Nica and Speicher [2006].

3.1. ***Algebraic model of a free arcsine pair.*** In Lecture Two we gave a group-theoretic construction of a pair of free random variables each of which has an arcsine distribution. In this example, the algebra of random variables is the group algebra $\mathscr{A} = \mathscr{A}[\mathbf{F}_2]$ of the free group on two generators A, B, and the expectation τ is the coefficient-of-id functional. We saw that the random variables

$$X = A + A^{-1}, \quad Y = B + B^{-1}$$

are freely independent, and each has an arcsine distribution:

$$\tau[X^n] = \tau[Y^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \binom{n}{n/2} & \text{if } n \text{ is even.} \end{cases}.$$

3.2. ***Algebraic model of a free semicircular pair.*** We can give a linear-algebraic model of a pair of free random variables each of which has a semicircular distribution. The ingredients in this construction are a complex vector space V and an inner product $B : \mathsf{V} \times \mathsf{V} \to \mathbb{C}$. Our random variables will be endomorphisms of the tensor algebra over V,

$$\mathfrak{F}(\mathsf{V}) = \bigoplus_{n=0}^{\infty} \mathsf{V}^{\otimes n},$$

which physicists and operator algebraists call the full Fock space over V after the Russian physicist Vladimir Fock. We view the zeroth tensor power $\mathsf{V}^{\otimes 0}$ as the line spanned by a distinguished unit vector $\mathsf{v}_\varnothing$ called the vacuum vector; $\mathsf{v}_\varnothing$ is an abstract vector which is not an element of V. Let $\mathcal{A} = \operatorname{End} \mathfrak{F}(V)$. This is a unital algebra, with unit the identity operator $I : \mathfrak{F}(V) \to \mathfrak{F}(V)$. To make $\mathcal{A}$ into a noncommutative probability space we need an expectation. We get an expectation by lifting the inner product on V to the inner product $\mathfrak{F}(B) : \mathfrak{F}(\mathsf{V}) \times \mathfrak{F}(\mathsf{V}) \to \mathbb{C}$ defined by

$$\mathfrak{F}(B)(\mathsf{v}_1 \otimes \cdots \otimes \mathsf{v}_m, \mathsf{w}_1 \otimes \cdots \otimes \mathsf{w}_n) = \delta_{mn} B(\mathsf{v}_1, \mathsf{w}_1) \dots B(\mathsf{v}_n, \mathsf{w}_n).$$

Note that this inner product makes $\mathcal{A} = \operatorname{End} \mathfrak{F}(B)$ into a $*$-algebra: for each $X \in \mathcal{A}$, X^* is that linear operator for which the equation

$$\mathfrak{F}(B)(X\mathsf{s}, \mathsf{t}) = \mathfrak{F}(B)(\mathsf{s}, X^*\mathsf{t})$$

holds true for every pair of tensors $\mathsf{s}, \mathsf{t} \in \mathfrak{F}(\mathsf{V})$. The expectation on $\mathcal{A}$ is the linear functional $\tau : \mathcal{A} \to \mathbb{C}$ defined by

$$\tau[X] = \mathfrak{F}(B)(X\mathsf{v}_\varnothing, \mathsf{v}_\varnothing).$$

This functional is called vacuum expectation. It is unital because

$$\tau[I] = \mathfrak{F}(B)(I\mathsf{v}_\varnothing, \mathsf{v}_\varnothing) = B(\mathsf{v}_\varnothing, \mathsf{v}_\varnothing) = 1.$$

Thus $(\mathcal{A}, \tau)$ is a noncommutative $*$-probability space.

To construct a semicircular element in $(\mathcal{A}, \tau)$, notice that to every nonzero vector $\mathsf{v} \in \mathsf{V}$ is naturally associated a pair of linear operators $R_\mathsf{v}, L_\mathsf{v} : \mathfrak{F}(V) \to \mathfrak{F}(V)$ whose action on decomposable tensors is defined by tensoring:

$$\begin{aligned} R_\mathsf{v}(\mathsf{v}_\varnothing) &= \mathsf{v}, \\ R_\mathsf{v}(\mathsf{v}_1 \otimes \cdots \otimes \mathsf{v}_n) &= \mathsf{v} \otimes \mathsf{v}_1 \otimes \cdots \otimes \mathsf{v}_n, \qquad n \geq 1, \end{aligned}$$

and insertion-contraction:

$$\begin{aligned} L_{\mathsf{v}}(\mathsf{v}_{\varnothing}) &= 0, \\ L_{\mathsf{v}}(\mathsf{v}_1) &= B(\mathsf{v}_1, \mathsf{v})\mathsf{v}_{\varnothing}, \\ L_{\mathsf{v}}(\mathsf{v}_1 \otimes \mathsf{v}_2 \otimes \cdots \otimes \mathsf{v}_n) &= B(\mathsf{v}_1, \mathsf{v})\mathsf{v}_2 \otimes \cdots \otimes \mathsf{v}_n, \quad n \geq 2. \end{aligned}$$

Since R_{v} maps $\mathsf{V}^{\otimes n} \to \mathsf{V}^{\otimes n+1}$ for each $n \geq 0$, it is called the raising (or creation) operator associated to v. Since L_{v} maps $\mathsf{V}^{\otimes n} \to \mathsf{V}^{\otimes n-1}$ for each $n \geq 1$ and kills the vacuum, it is called the lowering (or annihilation) operator associated to v. We have $R_{\mathsf{v}}^* = L_{\mathsf{v}}$, and also

$$L_{\mathsf{v}} R_{\mathsf{w}} = B(\mathsf{w}, \mathsf{v}) I$$

for any vectors $\mathsf{v}, \mathsf{w} \in \mathsf{V}$.

Let $\mathsf{v} \in \mathsf{V}$ be a unit vector, $B(\mathsf{v}, \mathsf{v}) = 1$, and consider the self-adjoint random variable

$$X_{\mathsf{v}} = L_{\mathsf{v}} + R_{\mathsf{v}}.$$

We claim that X_{v} has a semicircular distribution:

$$m_n(X_{\mathsf{v}}) = \tau[X_{\mathsf{v}}^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \mathrm{Cat}_{n/2} & \text{if } n \text{ is even.} \end{cases}$$

To see this, we write the expansion

$$\tau[X_{\mathsf{v}}^n] = \tau[(X_{\mathsf{v}} + Y_{\mathsf{v}})^n] = \sum_{W \in \{L_{\mathsf{v}}, R_{\mathsf{v}}\}^n} \tau[W],$$

where the summation is over all words of length n in the operators L_{v}, R_{v}. Only a very small fraction of these words have nonzero vacuum expectation. Using the relation $L_{\mathsf{v}} R_{\mathsf{v}} = I$ to remove occurrences of the substring $L_{\mathsf{v}} R_{\mathsf{v}}$, we see that any such word can be placed in normally ordered form

$$W = \underbrace{R_{\mathsf{v}} \dots R_{\mathsf{v}}}_{a} \underbrace{L_{\mathsf{v}} \dots L_{\mathsf{v}}}_{b}$$

with $a + b \leq n$. Since the lowering operator kills the vacuum vector, the vacuum expectation of W can only be nonzero if $b = 0$. On the other hand, since $\mathsf{V}^{\otimes a}$ is $\mathfrak{F}(B)$-orthogonal to $\mathsf{V}^{\otimes 0}$ for $a > 0$, we must also have $a = 0$ to obtain a nonzero contribution. Thus the only words which contribute to the above sum are those whose normally ordered form is that of the identity operator. If we replace each occurrence of L_{v} in W with a $+1$ and each occurrence of R_{v} in W with a -1, the condition that W reduces to I becomes the condition that the corresponding bitstring has total sum zero and all partial sums nonnegative. There are no such bitstrings for n odd, and as we saw in Lecture One when n is even the required bitstrings are counted by the Catalan number $\mathrm{Cat}_{n/2}$.

Now let V_1 and V_2 be B-orthogonal vector subspaces of V, each of dimension at least one, and choose unit vectors $\mathsf{x} \in \mathsf{V}_1, \mathsf{y} \in \mathsf{V}_2$. According to the above construction, the random variables

$$X = L_{\mathsf{x}} + R_{\mathsf{x}}, \quad Y = L_{\mathsf{y}} + R_{\mathsf{y}}$$

are semicircular. In fact, they are freely independent. To prove this, we must demonstrate that

$$\tau[f_1(X)g_1(Y)\dots f_k(X)g_k(Y)] = 0$$

whenever $f_1, g_1, \dots, f_k, g_k$ are polynomials such that

$$\tau[f_1(X)] = \tau[g_1(Y)] = \dots = \tau[f_k(X)] = \tau[g_k(Y)] = 0.$$

This hypothesis means that $f_i(X) = f_i(L_{\mathsf{x}} + R_{\mathsf{x}})$ is a polynomial in $L_{\mathsf{x}}, R_{\mathsf{x}}$ none of whose terms are words which normally order to I, and similarly $g_j(Y) = g_j(L_{\mathsf{y}} + R_{\mathsf{y}})$ is a polynomial in $L_{\mathsf{y}}, R_{\mathsf{y}}$ none of whose terms are words which normally order to I. Consequently, the alternating product

$$f_1(X)g_1(Y)\dots f_k(X)g_k(Y)$$

is a polynomial in the operators $L_{\mathsf{x}}, R_{\mathsf{x}}, L_{\mathsf{y}}, R_{\mathsf{y}}$ whose terms are words W of the form

$$W_{\mathsf{x}}^1 W_{\mathsf{y}}^1 \dots W_{\mathsf{x}}^k W_{\mathsf{y}}^k,$$

with W_{x}^i a word in $L_{\mathsf{x}}, R_{\mathsf{x}}$ which does not normally order to I and W_{y}^j a word in $L_{\mathsf{y}}, R_{\mathsf{y}}$ which does not normally order to I. Thus the only way that W can have a nonzero vacuum expectation is if we can use the relations $L_{\mathsf{x}}R_{\mathsf{y}} = B(\mathsf{y}, \mathsf{x})I$ and $L_{\mathsf{y}}R_{\mathsf{x}} = B(\mathsf{x}, \mathsf{y})I$ to normally order W as

$$B(\mathsf{x}, \mathsf{y})^m B(\mathsf{y}, \mathsf{x})^n I,$$

with m, n nonnegative integers at least one of which is positive. But, since x, y are B-orthogonal, this is the zero element of $\mathscr{A}$, which has vacuum expectation zero.

3.3. ***Algebraic versus asymptotic models.*** We have constructed algebraic models for a free arcsine pair and a free semicircular pair. Perhaps these should be called examples rather than models, since the term model connotes some degree of imprecision or ambiguity and algebra is a subject which allows neither.

Suppose that X, Y are free random variables living in an abstract noncommutative probability space $(\mathscr{A}, \tau)$. An approximate model for this pair will consist of a sequence $(\mathscr{A}_N, \tau_N)$ of concrete or canonical noncommutative probability spaces together with a sequence of pairs X_N, Y_N of random variables from these spaces such that X_N models X and Y_N models Y, i.e.,

$$\tau[f(X)] = \lim_{N\to\infty} \tau[f(X_N)], \quad \tau[g(Y)] = \lim_{N\to\infty} \tau[g(Y_N)]$$

for any polynomials f, g, and such that free independence holds in the large N limit, i.e.,

$$\lim_{N\to\infty} \tau[f_1(X_N)g_1(Y_N)\dots f_k(X_N)g_k(Y_N)] = 0$$

whenever $f_1, g_1, \dots, f_k, g_k$ are polynomials such that

$$\begin{aligned}\lim_{N\to\infty} \tau_N[f_1(X_N)] = \lim_{N\to\infty} \tau_N[g_1(Y_N)] = \cdots = \lim_{N\to\infty} \tau_N[f_k(X_N)] \\ = \lim_{N\to\infty} \tau_N[g_k(Y_N)] = 0.\end{aligned}$$

The question of which noncommutative probability spaces are considered concrete or canonical, and could therefore serve as potential models, is subjective and determined by individual experience. Three examples of concrete noncommutative probability spaces are:

Group probability spaces: $(\mathscr{A}, \tau)$ consists of the group algebra $\mathscr{A} = \mathscr{A}[\mathbf{G}]$ of a group $\mathbf{G}$, and τ is the coefficient-of-identity expectation. This noncommutative probability space is commutative if and only if $\mathbf{G}$ is abelian.

Classical probability spaces: $(\mathscr{A}, \tau)$ consists of the algebra of complex random variables $\mathscr{A} = L^{\infty-}(\Omega, \mathscr{F}, P) = \bigcap_{p=1}^{\infty} L^p(\Omega, \mathscr{F}, P)$ defined on a Kolmogorov triple which have finite absolute moments of all orders, and τ is the classical expectation $\tau[X] = \mathbb{E}[X]$. Classical probability spaces are always commutative.

Matrix probability spaces: $(\mathscr{A}, \tau)$ consists of the algebra $\mathscr{A} = \mathscr{M}_N(\mathbb{C})$ of $N \times N$ complex matrices $X = [X(ij)]$, and expectation is the normalized trace:

$$\tau[X] = \mathrm{tr}_N[X] = \frac{X(11) + \cdots + X(NN)}{N}.$$

This noncommutative probability space is commutative if and only if $N = 1$.

The first class of model noncommutative probability spaces, group probability spaces, is algebraic and we are trying to move away from algebraic examples. The second model class, classical probability spaces, has genuine randomness but is commutative. The third model class, matrix probability spaces, has a parameter N that can be pushed to infinity but has no randomness. By combining classical probability spaces and matrix probability spaces we arrive at a class of model noncommutative probability spaces which incorporate both randomness and a parameter which can be made large. Thus we are led to consider random matrices.

The space of $N \times N$ complex random matrices is the noncommutative probability space $(\mathcal{A}_N, \tau_N) = (L^{\infty-}(\Omega, \mathcal{F}, P) \otimes \mathcal{M}_N(\mathbb{C}), \mathbb{E} \otimes \mathrm{tr}_N)$. A random variable X_N in this space may be viewed as an $N \times N$ matrix whose entries $X_N(ij)$ belong to the algebra $L^{\infty-}(\Omega, \mathcal{F}, P)$. The expectation $\tau_N[X_N]$ is the expected value of the normalized trace:

$$\tau_N[X_N] = (\mathbb{E} \otimes \mathrm{tr}_N)[X_N] = \mathbb{E}\left[\frac{X_N(11) + \cdots + X_N(NN)}{N}\right].$$

We have already seen indications of a connection between free probability and random matrices. The fact that Wigner's semicircle law assumes the role of the Gaussian distribution in free probability signals a connection between these subjects. Another example is the occurrence of the Marchenko–Pastur distribution in the free version of the Poisson limit theorem — this distribution is well-known in random matrix theory in connection with the asymptotic eigenvalue distribution of Wishart matrices. In Lecture One, we were led to free independence when we tried to solve a counting problem associated to graphs drawn in the plane. The use of random matrices to enumerate planar graphs has been a subject of much interest in mathematical physics since the seminal work of Edouard Brézin, Claude Itzykson, Giorgio Parisi and Jean-Bernard Zuber [Brézin et al. 1978], which built on insights of Gerardus 't Hooft. Then, when we examined the dynamics of the semicircle flow, we found that the free analogue of the heat equation is the complex Burgers equation. This partial differential equation actually appeared in [Voiculescu 1986] before it emerged in random matrix theory [Matytsin 1994] and the discrete analogue of random matrix theory, the dimer model [Kenyon and Okounkov 2007].

In the remainder of these notes, we will model a pair of free random variables X, Y living in an abstract noncommutative probability space using sequences X_N, Y_N of random matrices living in random matrix space. This is first carried out in the important special case where X, Y are semicircular random variables, then adapted to allow Y to have arbitrary distribution while X remains semicircular, and finally relaxed to allow X, Y to have arbitrary specified distributions. The random matrix models of free random variables which we describe below were used by Voiculescu in order to resolve several previously intractable problems in the theory of von Neumann algebras; see [Mingo and Speicher $\geq$ 2014; Voiculescu et al. 1992] for more information. Random matrix models which approximate free random variables in a stronger sense than that described here were subsequently used by Uffe Haagerup and Steen Thorbjørnsen [2005] to resolve another operator algebras conjecture, this time concerning the Ext-invariant of the reduced C^*-algebra of $\mathbf{F}_2$. An important feature of the connection between free probability and random matrices is that it can sometimes be inverted to obtain information about random matrices using the free calculus. For each of the three matrix models constructed we give an example of this type.

3.4. ***Random matrix model of a free semicircular pair.*** In this subsection we construct a random matrix model for a free semicircular pair X, Y.

In Lecture One, we briefly discussed Wigner matrices. A real Wigner matrix is a symmetric matrix whose entries are centred real random variables which are independent up to the symmetry constraint. A complex Wigner matrix is a Hermitian matrix whose entries are centred complex random variables which are independent up to the complex symmetry constraint. Our matrix model for a free semicircular pair will be built out of complex Wigner matrices of a very special type: they will be GUE random matrices.

To construct a GUE random matrix X_N, we start with a Ginibre matrix Z_N. Let $(\Omega, \mathscr{F}, P)$ be a Kolmogorov triple. The N^2 matrix elements $Z_N(ij) \in L^{\infty-}(\Omega, \mathscr{F}, P)$ of a Ginibre matrix are iid complex Gaussian random variables of mean zero and variance $1/N$. Thus Z_N is a random variable in the noncommutative probability space $(\mathscr{A}_N, \tau_N) = (L^{\infty-}(\Omega, \mathscr{F}, P) \otimes \mathscr{M}_N(\mathbb{C}), \mathbb{E} \otimes \mathrm{tr}_N)$. The symmetrized random matrix $X_N = \frac{1}{2}(Z_N + Z_N^*)$ is again a member of random matrix space. The joint distribution of the eigenvalues of X_N can be explicitly computed, and is given by

$$P(\lambda_N(1) \in I_1, \dots, \lambda_N(N) \in I_N) \propto \int_{I_1} \dots \int_{I_N} e^{-N^2 \mathscr{H}(\lambda_1, \dots, \lambda_N)} \,\mathrm{d}\lambda_1 \dots \mathrm{d}\lambda_N$$

for any intervals $I_1, \dots, I_N \subseteq \mathbb{R}$, where $\mathscr{H}$ is the log-gas Hamiltonian [Forrester 2010]

$$\mathscr{H}(\lambda_1, \dots, \lambda_N) = \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_i^2}{2} - \frac{1}{N^2} \sum_{1 \le i \ne j \le N} \log |\lambda_i - \lambda_j|.$$

The random point process on the real line driven by this Hamiltonian is known as the Gaussian Unitary Ensemble, and X_N is termed a GUE random matrix. GUE random matrices sit at the nexus of the two principal strains of complex random matrix theory: they are simultaneously Hermitian Wigner matrices and unitarily invariant matrices. The latter condition means that the distribution of a GUE matrix in the space of $N \times N$ Hermitian matrices is invariant under conjugation by unitary matrices. The spectral statistics of a GUE random matrix can be computed in gory detail from knowledge of the joint distribution of eigenvalues, and virtually any question can be answered. The universality programme in random matrix theory seeks to show that, in the limit $N \to \infty$ and under mild hypotheses, Hermitian Wigner matrices as well as unitarily invariant Hermitian matrices exhibit the same spectral statistics as GUE matrices.

Given the central role of the GUE in random matrix theory, it is fitting that our matrix model for a free semicircular pair is built from a pair of independent GUE matrices. The first step in proving this is to show that a single GUE matrix X_N in random matrix space $(\mathcal{A}_N, \tau_N)$ is an asymptotic model for a single semicircular random variable X living in an abstract noncommutative probability space $(\mathcal{A}, \tau)$. In other words, we need to prove that

$$\lim_{N\to\infty} \tau_N[X_N^n] = \lim_{N\to\infty} (\mathbb{E}\otimes \mathrm{tr}_N)[X_N^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \mathrm{Cat}_{n/2} & \text{if } n \text{ is even.} \end{cases}$$

In order to establish this, we will not need access to the eigenvalues of X_N. Rather, we work with the correlation functions of its entries.

Let $X_N = [X_N(ij)]$ be a GUE random matrix. Mixed moments of the random variables $X_N(ij)$, i.e., expectations of the form

$$\mathbb{E}\Bigg[\prod_{k=1}^{n} X_N(i(k)j(k))\Bigg]$$

where i, j are functions $[n] \to [N]$, are called correlation functions. All correlations may be computed in terms of pair correlations (i.e., covariances)

$$\mathbb{E}[X_N(ij)\overline{X_N(kl)}] = \mathbb{E}[X_N(ij)X_N(lk)] = \frac{\delta_{ik}\delta_{jl}}{N}$$

using a convenient combinatorial formula known as Wick's formula. This formula, named for the Italian physicist Gian-Carlo Wick, is yet another manifestation of the moment-cumulant/exponential formulas. It asserts that

$$\mathbb{E}\Bigg[\prod_{k=1}^{n} X_N(i(k)j(k)))\Bigg] = \sum_{\pi\in \mathsf{P}_2(n)} \prod_{\{r,s\}\in\pi} \mathbb{E}[X_N(i(r)j(r))X_N(i(s)j(s))]$$

for any integer $n \geq 1$ and functions $i, j : [n] \to [N]$. The sum on the right hand side is taken over all pair partitions of $[n]$, and the product is over the blocks of π. For example,

$$\mathbb{E}[X_N(i(1)j(1))X_N(i(2)j(2))X_N(i(3)j(3))] = 0$$

since there are no pairings on three points, whereas

$$\begin{aligned} &\mathbb{E}[X_N(i(1)j(1))X_N(i(2)j(2))X_N(i(3)j(3))X_N(i(4)j(4))] \\ &\qquad = \mathbb{E}[X_N(i(1)j(1))X_N(i(2)j(2))]\,\mathbb{E}[X_N(i(3)j(3))X_N(i(4)j(4))] \\ &\qquad\quad + \mathbb{E}[X_N(i(1)j(1))X_N(i(3)j(3))]\,\mathbb{E}[X_N(i(2)j(2))X_N(i(4)j(4))] \\ &\qquad\quad + \mathbb{E}[X_N(i(1)j(1))X_N(i(4)j(4))]\,\mathbb{E}[X_N(i(2)j(2))X_N(i(3)j(3))], \end{aligned}$$

corresponding to the three pair partitions

$$\{1,2\}\sqcup\{3,4\},\quad \{1,3\}\sqcup\{2,4\},\quad \{1,4\}\sqcup\{2,3\}$$

of [4]. The Wick formula is a special feature of Gaussian random variables which, ultimately, is a consequence of the moment formula

$$\mathbb{E}[X^n]=\sum_{\pi\in\mathrm{P}_2(n)}1$$

for a single standard real Gaussian X which we proved in Lecture One. A proof of the Wick formula may be found in Alexandre Zvonkin's article [1997].

We now compute the moments of the trace of a GUE matrix X_N using the Wick formula, and then take the $N\to\infty$ limit. We have

$$\begin{aligned}\tau_N[X_N^n]&=\frac{1}{N}\sum_{i:[n]\to[N]}\mathbb{E}[X_N(i(1)i(2))X_N(i(2)i(3)))\dots X_N(i(n)i(1))]\\&=\frac{1}{N}\sum_{i:[n]\to[N]}\mathbb{E}\biggl[\prod_{k=1}^{n}X_N(i(k)i\gamma(k))\biggr],\end{aligned}$$

where $\gamma=(1\;2\;\dots\;n)$ is the full forward cycle in the symmetric group $\mathbf{S}(n)$. Let us apply the Wick formula to each term of this sum, and then use the covariance structure of the matrix elements. We obtain

$$\begin{aligned}\mathbb{E}\biggl[\prod_{k=1}^{n}X_N(i(k)i\gamma(k))\biggr]&=\sum_{\pi\in\mathrm{P}_2(n)}\prod_{\{r,s\}\in\pi}\mathbb{E}[X_N(i(r)i\gamma(r))X_N(i(s)i\gamma(s))]\\&=N^{-n/2}\sum_{\pi\in\mathrm{P}_2(n)}\prod_{\{r,s\}\in\pi}\delta_{i(r)i\gamma(s)}\delta_{i(s)i\gamma(r)}.\end{aligned}$$

Now, any pair partition of $[n]$ can be viewed as a product of disjoint two-cycles in $\mathbf{S}(n)$. For example, the three pair partitions of [4] enumerated above may be viewed as the fixed-point-free involutions

$$(1\;2)(3\;4),\;\;(1\;3)(2\;4),\;\;(1\;4)(2\;3)$$

in $\mathbf{S}(4)$. This is a useful shift in perspective because partitions are inert combinatorial objects whereas permutations are functions which act on points. Our computation above may thus be rewritten as

$$\mathbb{E}\biggl[\prod_{k=1}^{n}X_N(i(k)i\gamma(k))\biggr]=N^{-n/2}\sum_{\pi\in\mathrm{P}_2(n)}\prod_{k=1}^{n}\delta_{i(k)i\gamma\pi(k)}.$$

Putting this all together and changing order of summation, we obtain

$$\begin{aligned}\tau_N[X_N^n] &= N^{1-n/2} \sum_{i:[n]\to[N]} \sum_{\pi\in \mathsf{P}_2(n)} \prod_{k=1}^{n} \delta_{i(k)i\gamma\pi(k)} \\ &= N^{1-n/2} \sum_{\pi\in \mathsf{P}_2(n)} \sum_{i:[n]\to[N]} \prod_{k=1}^{n} \delta_{i(k)i\gamma\pi(k)},\end{aligned}$$

from which we see that the internal sum is nonzero if and only if the function $i:[n]\to[N]$ is constant on the cycles of the permutation $\gamma\pi\in\mathbf{S}(n)$. In order to build such a function, we must specify one of N possible values to be taken on each cycle. We thus obtain

$$\tau_N[X_N^n] = \sum_{\pi\in \mathsf{P}_2(n)} N^{c(\gamma\pi)-1-n/2},$$

where $c(\sigma)$ denotes the number of cycles in the disjoint cycle decomposition of a permutation $\sigma\in\mathbf{S}(n)$. For example, when $n=3$ we have $\tau_n[X_N^3]=0$ since there are no fixed-point-free involutions in $\mathbf{S}(3)$. In order to compute $\tau_N[X_N^4]$, we first compute the product of γ with all fixed-point-free involutions in $\mathbf{S}(4)$,

$$\begin{aligned}(1\ 2\ 3\ 4)(1\ 2)(3\ 4) &= (1\ 3)(2)(4)\\ (1\ 2\ 3\ 4)(1\ 3)(2\ 4) &= (1\ 4\ 3\ 2)\\ (1\ 2\ 3\ 4)(1\ 4)(2\ 3) &= (2\ 4)(1)(3),\end{aligned}$$

and from this we obtain

$$\tau_N[X_N^4] = 2+\frac{1}{N^2}.$$

More generally, $\tau_N[X_N^n]=0$ whenever n is odd since there are no pairings on an odd number of points. When $n=2k$ is even the product $\gamma\pi$ has the form

$$\gamma\pi = (1\ 2\ \ldots\ 2k)(s_1\ t_1)(s_2\ t_2)\ldots(s_k\ t_k).$$

In this product, each transposition factor $(s_i\ t_i)$ acts either as a “cut” or as a “join”, meaning that it may either cut a cycle of $(1\ 2\ \ldots\ 2k)(s_1\ t_1)\ldots(s_{i-1}\ t_{i-1})$ in two, or join two disjoint cycles together into one. More geometrically, we can view the product $\gamma\pi$ as a walk of length k on the (right) Cayley graph of $\mathbf{S}(2k)$; this walk is nonbacktracking and each step taken augments the distance from the identity permutation by ±1 (see Figure 15).

A cut (step towards the identity) occurs when s_i and t_i reside on the same cycle in the disjoint cycle decomposition of $(1\ 2\ \ldots\ 2k)(s_1\ t_1)\ldots(s_{i-1}\ t_{i-1})$, while a join (step away from the identity) occurs when s_i and t_i are on different cycles. In general, the number of cycles in the product will be

$$c(\gamma\pi) = 1+\#\text{cuts}-\#\text{joins},$$

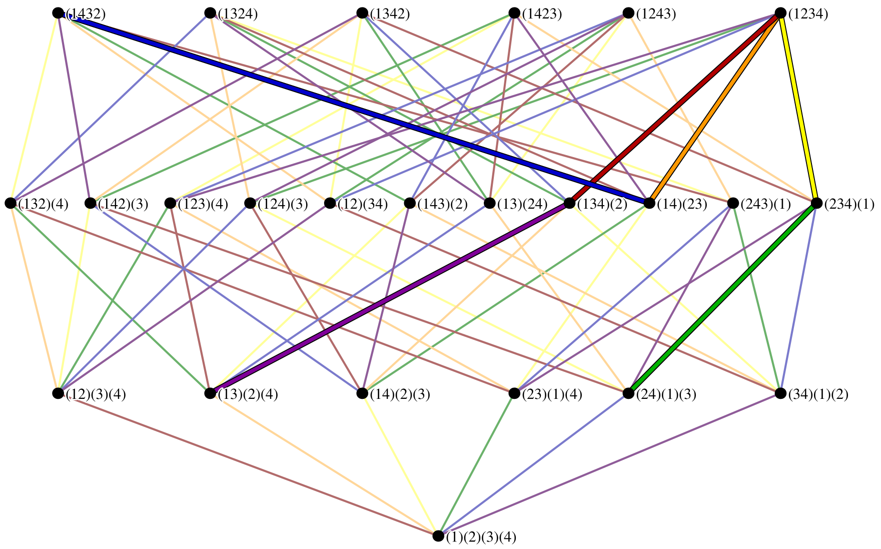

Figure 15. Walks corresponding to the products $\gamma\pi$ in $\mathbf{S}(4)$.

so $c(\gamma\pi)$ is maximal at $c(\gamma\pi) = 1+k$ when it is acted on by a sequence of k cut transpositions. In this case we get a contribution of $N^{1+k-1-k} = N^0$ to $\tau[X_N^n]$. In fact, we always have

$$\#\text{cuts} - \#\text{joins} = k - 2g$$

for some nonnegative integer g, leading to a contribution of the form N^{-2g} and resulting in the formula

$$\tau_N[X_N^{2k}] = \sum_{g\geq 0} \frac{\varepsilon_g(2k)}{N^{2g}}$$

where $\varepsilon_g(2k)$ is the number of products $\gamma\pi$ of the long cycle with a fixed-point-free involution in $\mathbf{S}(2k)$ which terminate at a point of the sphere $\partial B(\text{id}, 2k-1-2g)$. We are only interested in the first term of this expansion, $\varepsilon_0(2k)$, which counts fixed-point-free involutions in $\mathbf{S}(2k)$ entirely composed of cuts. It is not difficult to see that $(s_1\ t_1)\dots(s_k\ t_k)$ is a sequence of cuts for γ if and only if it corresponds to a noncrossing pair partition of $[2k]$, and as we know the number of these is Cat_k.

We have now shown that

$$\lim_{N\to\infty} \tau_N[X_N^n] = \lim_{N\to\infty} (\mathbb{E}\otimes\text{tr}_N)[X_N^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \text{Cat}_{n/2} & \text{if } n \text{ is even.} \end{cases}$$

for a GUE matrix X_N. This establishes that X_N is an asymptotic random matrix model of a single semicircular random variable X. It remains to use this fact to construct a sequence of pairs of random matrices which model a pair X, Y of freely independent semicircular random variables.

What should we be looking for? Let $X^{(1)}$, $X^{(2)}$ be a pair of free semicircular random variables. Let $e : [n] \to [2]$ be a function, and apply the free moment-cumulant formula to the corresponding mixed moment:

$$\begin{aligned}\tau[X^{(e(1))} \dots X^{(e(n))}] &= \sum_{\pi \in \mathsf{NC}(n)} \prod_{B \in \pi} \kappa_{|B|}(X^{(e(i))} : i \in B) \\ &= \sum_{\pi \in \mathsf{NC}_2(n)} \prod_{\{r,s\} \in \pi} \delta_{e(r)e(s)}.\end{aligned}$$

This reduction occurs because $X^{(1)}$, $X^{(2)}$ are free, so that all mixed free cumulants in these variables vanish. Moreover, these variables are semicircular so only order two pure cumulants survive. We can think of the function e as a bicolouring of $[n]$. The formula for mixed moments of a semicircular pair then becomes

$$\tau[X^{(e(1))} \dots X^{(e(n))}] = \sum_{\pi \in \mathsf{NC}_2^{(e)}(n)} 1,$$

where $\pi \in \mathsf{NC}_2^{(e)}(n)$ is the set of noncrossing pair partitions of $[n]$ which pair elements of the same colour. This is very much like the Wick formula for Gaussian expectations, but with Gaussians replaced by semicirculars and summation restricted to noncrossing pairings. We need to realize this structure in the combinatorics of GUE random matrices.

This construction goes as follows. Let $Z_N^{(e)}(ij),\ 1 \le e \le 2,\ 1 \le i, j \le N$ be a collection of $2N^2$ iid centred complex Gaussian random variables of variance $1/N$. Form the corresponding Ginibre matrices $Z_N^{(1)} = [Z_N^{(1)}(ij)],\ Z_N^{(2)} = [Z_N^{(2)}(ij)]$ and GUE matrices $X_N^{(1)} = \frac{1}{2}(Z_N^{(1)} + (Z_N^{(1)})^*),\ X_N^{(2)} = \frac{1}{2}(Z_N^{(2)} + (Z_N^{(2)})^*)$. The resulting covariance structure of matrix elements is

$$\mathbb{E}[X_N^{(p)}(ij)\overline{X_N^{(q)}(kl)}] = \mathbb{E}[X_N^{(p)}(ij)X_N^{(q)}(lk)] = \frac{\delta_{ik}\delta_{jl}\delta_{pq}}{N}.$$

We can prove that $X_N^{(1)}$, $X_N^{(2)}$ are asymptotically free by showing that

$$\lim_{N \to \infty} \tau_N[X_N^{(e(1))} \dots X_N^{(e(n))}] = |\mathsf{NC}_2^{(e)}(n)|,$$

and this can in turn be proved using the Wick formula and the above covariance structure. Computations almost exactly like those appearing in the one-matrix case lead to the formula

$$\tau_N[X_N^{(e(1))} \dots X_N^{(e(n))}] = \sum_{\pi \in \mathsf{P}_2^{(e)}(n)} N^{c(\gamma\pi)-1-n/2},$$

with the summation being taken over the set $\mathsf{P}_2^{(e)}(n)$ of pairings on $[n]$ which respect the colouring $e : [n] \to [2]$. Arguing as above, each such pairing makes a contribution of the form N^{-2g} for some $g \geq 0$, and those which make contributions on the leading order N^0 correspond to sequences of cut transpositions for the full forward cycle π, which we know come from noncrossing pairings. So in the limit $N \to \infty$ this expectation converges to $|\mathsf{NC}_2^{(e)}(n)|$, as required.

3.5. ***Random matrix model of a free pair with one semicircle.*** In the previous subsection we modelled a free pair of semicircular random variables X, Y living in an abstract noncommutative probability space $(\mathscr{A}, \tau)$ using a sequence of independent GUE random matrices X_N, Y_N living in random matrix space $(\mathscr{A}_N, \tau_N)$.

It is reasonable to wonder whether we have not overlooked the possibility of modelling X, Y in a simpler way, namely using deterministic matrices. Indeed, we have

$$\tau[X^n] = \int_{\mathbb{R}} t^n \mu_X(\mathrm{d}t)$$

with

$$\mu_X(\mathrm{d}t) = \frac{1}{2\pi}\sqrt{4-t^2}\,\mathrm{d}t$$

the Wigner semicircle measure, and this fact leads to a deterministic matrix model for X. For each $N \geq 1$, define the N-th classical locations $L_N(1) < L_N(2) < \cdots < L_N(N)$ of μ_X implicitly by

$$\int_{-2}^{L_N(i)} \mu_X(\mathrm{d}t) = \frac{i}{N}.$$

That is, we start at $t = -2$ and integrate along the semicircle until a mass of i/N is achieved, at which time we mark off the corresponding location $L_N(i)$ on the t-axis. The measure μ_N which places mass $1/N$ at each of the N-th classical locations converges weakly to μ_X as $N \to \infty$. Consequently, the diagonal matrix X_N with entries $X_N(ij) = \delta_{ij} L_N(i)$ is a random variable in deterministic matrix space $(\mathscr{M}_N(\mathbb{C}), \mathrm{tr}_N)$ which models X,

$$\lim_{N\to\infty} \mathrm{tr}_N[X_N^n] = \tau[X^n].$$

Since X and Y are equidistributed, putting $Y_N := X_N$ we have that X_N models X and Y_N models Y. However, X_N and Y_N are not asymptotically free. Indeed, asymptotic freeness of X_N and Y_N would imply that

$$\lim_{N\to\infty} \mathrm{tr}_N[X_N Y_N] = \lim_{N\to\infty} \mathrm{tr}_N[X_N] \lim_{N\to\infty} \mathrm{tr}_N[X_N] = 0,$$

but instead we have

$$\mathrm{tr}_N[X_N Y_N] = \frac{L_N(1)^2 + \cdots + L_N(N)^2}{N},$$

the mean squared classical locations of the Wigner measure, which is strictly positive and increasing in N. Thus while X_N and Y_N model X and Y respectively, they cannot model the free relation between them. However, this does not preclude the possibility that a pair of free random variables can be modelled by one random and one deterministic matrix.

Let X and Y be a pair of free random variables with X semicircular, and Y of arbitrary distribution. Let X_N be a sequence of GUE matrices modelling X, and suppose that Y_N is a sequence of deterministic matrices modelling Y,

$$\lim_{N\to\infty} \mathrm{tr}_N[Y_N^n] = \tau[Y^n].$$

X_N lives in random matrix space $(\mathscr{A}_N, \tau_N) = (L^{\infty-}(\Omega, \mathscr{F}, P) \otimes \mathscr{M}_N(\mathbb{C}), \mathbb{E} \otimes \mathrm{tr}_N)$ while Y_N lives in deterministic matrix space $(\mathscr{M}_N(\mathbb{C}), \mathrm{tr}_N)$, so a priori it is meaningless to speak of the potential asymptotic free independence of X_N and Y_N. However, we may think of a deterministic matrix as a random matrix whose entries are constant random variables in $L^{\infty-}(\Omega, \mathscr{F}, P)$. This corresponds to an embedding of deterministic matrix space in random matrix space satisfying $\tau_N|_{\mathscr{M}_N(\mathbb{C})} = (\mathbb{E} \otimes \mathrm{tr}_N)|_{\mathscr{M}_N(\mathbb{C})} = \mathrm{tr}_N$. From this point of view, Y_N is a random matrix model of Y and we can consider the possibility that $X_N, Y_N \in \mathscr{A}_N$ are asymptotically free with respect to τ_N. We now show that this is indeed the case.

As in the previous subsection, we proceed by identifying the combinatorial structure governing the target pair X, Y and then looking for this same structure in the $N \to \infty$ asymptotics of X_N, Y_N. Our target is a pair of free random variables with X semicircular and Y arbitrary. Understanding their joint distribution means understanding the collection of mixed moments

$$\tau[X^{p(1)} Y^{q(1)} \ldots X^{p(n)} Y^{q(n)}],$$

with $n \geq 1$ and $p, q : [n] \to \{0, 1, 2, \ldots\}$. This amounts to understanding mixed moments of the form

$$\tau[X Y^{q(1)} \ldots X Y^{q(n)}],$$

since we can artificially insert copies of $Y^0 = 1_{\mathscr{A}}$ to break up powers of X greater than one. We can expand this expectation using the free moment-cumulant formula and simplify the resulting expression using the fact that mixed cumulants in free random variables vanish. Further simplification results from the fact that, since X is semicircular, its only nonvanishing pure cumulant is $\kappa_2(X) = 1$. This leads to a formula for $\tau[X Y^{q(1)} \ldots X Y^{q(n)}]$ which is straightforward but whose statement requires some notions which we have not covered (in particular, the complement of a noncrossing partition; see [Nica and Speicher 2006]). However, in the case where τ is a tracial expectation, meaning that $\tau[AB]=\tau[BA]$, the formula in question can be stated more simply as

$$\tau[XY^{q(1)}\dots XY^{q(n)}]=\sum_{\pi\in\mathsf{NC}_2(n)}\tau_{\pi\gamma}[Y^{q(1)},\dots,Y^{q(n)}].$$

Here, as in the last subsection, we think of a pair partition $\pi\in\mathsf{P}_2(n)$ as a product of disjoint two-cycles in the symmetric group $\mathbf{S}(n)$, and γ is the full forward cycle $(1\ 2\ \dots\ n)$. Given a permutation $\sigma\in\mathbf{S(n)}$, the expression $\tau_\sigma[A_1,\dots,A_N]$ is defined to be the product of τ extended over the cycles of σ. For example,

$$\tau_{(1\ 6\ 2)(4\ 5)(3)}[A_1,A_2,A_3,A_4,A_5,A_6]=\tau[A_1A_6A_2]\tau[A_4A_5]\tau[A_3].$$

This definition is kosher since τ is tracial. We now have our proof strategy: we will prove that X_N,Y_N are asymptotically free by showing that

$$\lim_{N\to\infty}\tau_N[X_NY_N^{q(1)}\dots X_NY_N^{q(n)}]=\sum_{\pi\in\mathsf{NC}_2(n)}\tau_{\pi\gamma}[Y^{q(1)},\dots,Y^{q(n)}].$$

The computation proceeds much as in the last section — we expand everything in sight and apply the Wick formula. We have

$$\begin{aligned}\tau_N[X_NY_N^{q(1)}\dots X_NY_N^{q(n)}]&\\ &=\frac{1}{N}\sum_a\mathbb{E}\big[X_N(a(1)a(2))Y_N^{q(1)}(a(2)a(3))\cdots\\ &\qquad\times X_N(a(2n-1)a(2n))Y_N^{q(n)}(a(2n)a(1))\big],\end{aligned}$$

the summation being over all functions $a:[2n]\to[N]$. Let us reparametrize each term of the sum with $i,j:[n]\to[N]$ defined by

$$(a(1),a(2),\dots,a(2n-1),a(2n))=(i(1),j(1),\dots,i(n),j(n)).$$

Our computation so far becomes

$$\begin{aligned}\tau_N[X_NY_N^{q(1)}\dots X_NY_N^{q(n)}]&\\ &=\frac{1}{N}\sum_{i,j}\mathbb{E}\Bigg[\prod_{k=1}^nX_N(i(k)j(k))\Bigg]\prod_{k=1}^nY_N^{q(k)}(j(k)i\gamma(k)).\end{aligned}$$

Applying the Wick formula, the calculation evolves as follows:

$$\begin{aligned}
&\tau_N[X_N Y_N^{q(1)} \dots X_N Y_N^{q(n)}] \\
&\quad = \frac{1}{N} \sum_{i,j} \sum_{\pi \in \mathsf{P}_2(n)} \prod_{\{r,s\} \in \pi} \mathbb{E}[X_N(i(r)j(r)) X_N(i(s)j(s))] \prod_{k=1}^{n} Y_N^{q(k)}(j(k) i\gamma(k)) \\
&\quad = N^{-1-n/2} \sum_{i,j} \sum_{\pi \in \mathsf{P}_2(n)} \prod_{k=1}^{n} \delta_{i(k) j\pi(k)} Y_N^{q(k)}(j(k) i\gamma(k)) \\
&\quad = N^{-1-n/2} \sum_{\pi \in \mathsf{P}_2(n)} \sum_{j} \prod_{k=1}^{n} Y_N^{q(k)}(j(k) j\pi\gamma(k)) \\
&\quad = N^{-1-n/2} \sum_{\pi \in \mathsf{P}_2(n)} \mathrm{Tr}_{\pi\gamma}[Y_N^{q(1)}, \dots, Y_N^{q(n)}] \\
&\quad = \sum_{\pi \in \mathsf{P}_2(n)} N^{c(\pi\gamma)-1-n/2} \,\mathrm{tr}_{\pi\gamma}[Y_N^{q(1)}, \dots, Y_N^{q(n)}].
\end{aligned}$$

As in the previous subsection, the dominant contributions to this sum are of order N^0 and come from those pair partitions $\pi \in \mathsf{P}_2(n)$ for which $c(\pi\gamma)$ is maximal, and these are the noncrossing pairings. Hence we obtain

$$\lim_{N\to\infty} \tau_N[X_N Y_N^{q(1)} \dots X_N Y_N^{q(n)}] = \sum_{\pi \in \mathsf{NC}_2(n)} \tau_{\pi\gamma}[Y^{q(1)}, \dots, Y^{q(n)}],$$

as required.

3.6. ***Random matrix model of an arbitrary free pair.*** In the last section we saw that a pair of free random variables can be modelled by one random and one deterministic matrix provided that at least one of the target variables is semicircular. In this case, the semicircular target is modelled by a sequence of GUE random matrices.

In this section we show that any pair of free random variables can be modelled by one random and one deterministic matrix, provided each target variable can be individually modelled by a sequence of deterministic matrices. The idea is to randomly rotate one of the deterministic matrix models so as to create the free relation.

Let X, Y be a pair of free random variables living in an abstract noncommutative probability space $(\mathcal{A}, \tau)$. We make no assumption on their moments. What we assume is the existence of a pair of deterministic matrix models

$$\tau[X^n] = \lim_{N\to\infty} \mathrm{tr}_N[X_N^n], \qquad \tau[Y^n] = \lim_{N\to\infty} \mathrm{tr}_N[Y_N^n].$$

If X, Y happen to have distributions μ_X, μ_Y which are compactly supported probability measures on $\mathbb{R}$, then such models can always be constructed. In particular, this will be the case if X, Y are bounded self-adjoint random variables living in a $*$-probability space.

As in the previous subsection, we view X_N, Y_N as random matrices with constant entries so that they reside in random matrix space $(\mathcal{A}_N, \tau_N)$, with the $\mathbb{E}$ part of $\tau_N = \mathbb{E} \otimes \mathrm{tr}_N$ acting trivially. As we saw above, there is no guarantee that X_N, Y_N are asymptotically free. On the other hand, we also saw that special pairs of free random variables can be modelled by one random and one deterministic matrix. Therefore it is reasonable to hope that making X_N genuinely random might lead to asymptotic freeness. We have to randomize X_N in such a way that its moments will be preserved. This can be achieved via conjugation by a unitary random matrix $U_N \in \mathcal{A}_N$,

$$X_N \mapsto U_N X_N U_N^*.$$

The deterministic matrix X_N and its randomized version $U_N X_N U_N^*$ have the same moments since

$$\begin{aligned}\tau_N[(U_N X_N U_N^*)^n] &= (\mathbb{E} \otimes \mathrm{tr}_N)[(U_N X_N U_N^*)^n] \\ &= (\mathbb{E} \otimes \mathrm{tr}_N)[U_N X_N^n U_N^*] \\ &= (\mathbb{E} \otimes \mathrm{tr}_N)[U_N^* U_N X_N^n] \\ &= (\mathbb{E} \otimes \mathrm{tr}_N)[X_N^n] \\ &= \tau_N[X_N^n].\end{aligned}$$

Consequently, the sequence $U_N X_N U_N^*$ is a random matrix model for X.

We aim to prove that $U_N X_N U_N^*$ and Y_N are asymptotically free. Since we are making no assumptions on the limiting variables X, Y, we cannot verify this by looking for special structure in the limiting mixed moments of $U_N X_N U_N^*$ and Y_N, as we did above. Instead, we must verify asymptotic freeness directly, using the definition:

$$\lim_{N\to\infty} \tau_N[f_1(U_N X_N U_N^*) g_1(Y_N) \dots f_n(U_N X_N U_N^*) g_n(Y_N)] = 0$$

whenever $f_1, g_1, \dots, f_n, g_n$ are polynomials such that

$$\begin{aligned}\lim_{N\to\infty} \tau_N[f_1(U_N X_N U_N^*)] &= \lim_{N\to\infty} \tau_n[g_1(Y_N)] = \cdots \\ &= \lim_{N\to\infty} \tau_N[f_n(U_N X_N U_N^*)] = \lim_{N\to\infty} \tau_n[g_n(Y_N)] = 0.\end{aligned}$$

Though the brute force verification of this criterion may seem an impossible task, we will see that it can be accomplished for a well-chosen sequence of unitary random matrices U_N. Let us advance as far as possible before specifying U_N precisely.

As an initial reduction, note the identity

$$\tau_N[f_1(U_N X_N U_N^*)g_1(Y_N)\dots f_n(U_N X_N U_N^*)g_n(Y_N)] \\ = \tau_N[U_N f_1(X_N)U_N^* g_1(Y_N)\dots U_N f_n(X_N)U_N^* g_n(Y_N)].$$

Since the f_i and g_j are polynomials and τ_N is linear, the right hand side of this equation may be expanded as a sum of monomial expectations,

$$\tau_N[U_N f_1(X_N)U_N^* g_1(Y_N)\dots U_N f_n(X_N)U_N^* g_n(Y_N)] \\ = \sum_{p,q} c(pq)\tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)}\dots U_N X_N^{p(n)} U_N^* Y_N^{q(n)}]$$

weighted by some scalar coefficients $c(pq)$, the sum being over functions $p : [n] \to \{0,\dots,\max\deg f_i\}$, $q : [n] \to \{0,\dots,\max\deg g_j\}$. Each monomial expectation can in turn be expanded as

$$\begin{aligned}\tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)}\dots U_N X_N^{p(n)} U_N^* Y_N^{q(n)}] \\ &= \frac{1}{N}\sum_a \mathbb{E}\big[U_N(a(1)a(2))X_N^{p(1)}(a(2)a(3))\dots \\ &\qquad \times U_N^*(a(4n-1)a(4n))Y_N^{q(n)}(a(4n)a(1))\big] \\ &= \frac{1}{N}\sum_a \mathbb{E}\big[U_N(a(1)a(2))X_N^{p(1)}(a(2)a(3))\dots \\ &\qquad \times \overline{U}_N(a(4n)a(4n-1))Y_N^{q(n)}(a(4n)a(1))\big].\end{aligned}$$

Let us reparametrize the summation index $a : [4n] \to [N]$ by a quadruple of functions $i, j, i', j' : [n] \to [N]$ according to

$$(a(1), a(2), a(3), a(4), \dots, a(4n-3), a(4n-2), a(4n-1), a(4n)) \\ = (i(1), j(1), j'(1), i'(1), \dots, i(n), j(n), j'(n), i'(n)).$$

Our monomial expectations then take the more streamlined form

$$\begin{aligned}\tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)}\dots U_N X_N^{p(n)} U_N^* Y_N^{q(n)}] \\ &= \frac{1}{N}\sum_{i,j,i',j'} \mathbb{E}\Bigg[\prod_{k=1}^n U_N(i(k)j(k))\overline{U}_N(i'(k)j'(k))\Bigg] \\ &\qquad \times \prod_{k=1}^n X_N^{p(k)}(j(k)j'(k))Y_N^{q(k)}(i'(k)i\gamma(k)),\end{aligned}$$

where as always $\gamma = (1\ 2\ \dots\ n)$ is the full forward cycle in the symmetric group $\mathbf{S}(n)$. In order to go any further with this calculation, we must deal with the correlation functions

$$\mathbb{E}\left[\prod_{k=1}^{n} U_N(i(k)j(k))\overline{U}_N(i'(k)j'(k))\right].$$

of the matrix elements of U_N. We would like to have an analogue of the Wick formula which will enable us to address these correlation functions. A formula of this type is known for random matrices sampled from the Haar probability measure on the unitary group $\mathbf{U}(N)$.

Haar-distributed unitary matrices are the second most important class of random matrices after GUE matrices. Like GUE matrices, they can be constructively obtained from Ginibre matrices. Let $\tilde{Z}_N = \sqrt{N} Z_N$ be an $N \times N$ random matrix whose entries $\tilde{Z}_N(ij)$ are iid complex Gaussian random variables of mean zero and variance one. This is a renormalized version of the Ginibre matrix which we previously used to construct a GUE random matrix. The Ginibre matrix $\tilde{Z}_N$ is almost surely nonsingular. Applying the Gram–Schmidt orthonormalization procedure to the columns of $\tilde{Z}_N$, we obtain a random unitary matrix U_N whose distribution in the unitary group $\mathbf{U}(N)$ is given by the Haar probability measure. The entries $U_N(ij)$ are bounded random variables, so U_N is a noncommutative random variable living in random matrix space $(\mathcal{A}_N, \tau_N)$. The eigenvalues $\lambda_N(1) = e^{\mathbf{i}\theta_N(1)}, \dots, \lambda_N(N) = e^{\mathbf{i}\theta_N(N)}$, $0 \leq \theta_N(1) \leq \dots \leq \theta_N(N) \leq 2\pi$ of U_N form a random point process on the unit circle with joint distribution

$$P(\theta_N(1) \in I_1, \dots, \theta_N(N) \in I_N) \propto \int_{I_1} \dots \int_{I_N} e^{-N^2 \mathcal{H}(\theta_1, \dots, \theta_N)} \, \mathrm{d}\theta_1 \dots \mathrm{d}\theta_N$$

for any intervals $I_1, \dots, I_N \subseteq [0, 2\pi]$, where $\mathcal{H}$ is the log-gas Hamiltonian [Forrester 2010]

$$\mathcal{H}(\theta_1, \dots, \theta_N) = -\frac{1}{N^2} \sum_{1 \leq i \neq j \leq N} \log |e^{\mathbf{i}\theta_i} - e^{\mathbf{i}\theta_j}|.$$

The random point process on the unit circle driven by this Hamiltonian is known as the Circular Unitary Ensemble, and U_N is termed a CUE random matrix. As with GUE random matrices, almost any question about the spectrum of CUE random matrices can be answered using this explicit formula; see, for example, [Diaconis 2003] for a survey of many interesting results.

We are not interested in the eigenvalues of CUE matrices, but rather in the correlation functions of their matrix elements. These can be handled using a Wick-type formula known as the Weingarten formula, after the American physicist Donald H. Weingarten.[3] Like the Wick formula, the Weingarten formula is a combinatorial rule which reduces the computation of general correlation

[3] Further information on Weingarten and his colleagues in the first Fermilab theory group may be found at http://bama.ua.edu/~lclavell/Weston/.

functions to the computation of a special class of correlations. Unfortunately, the Weingarten formula is more complicated than the Wick formula. It reads

$$\mathbb{E}\biggl[\prod_{k=1}^{n} U_N(i(k)j(k))\bar{U}_N(i'(k)j'(k))\biggr] = \sum_{\rho,\sigma\in\mathbf{S}(n)} \delta_{i\sigma,i'}\delta_{j\rho,j'}\mathbb{E}\biggl[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\rho^{-1}\sigma(k))\biggr].$$

Note that his formula only makes sense when $N \geq n$, and instead of a sum over fixed-point-free involutions we are faced with a double sum over all of $\mathbf{S}(n)$. Worse still, the Weingarten formula does not reduce our problem to the computation of pair correlators, but only to the computation of arbitrary permutation correlators

$$\mathbb{E}\biggl[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\pi(k))\biggr], \quad \pi \in \mathbf{S}(n),$$

and these have a rather complicated structure. Their computation is the subject of a large literature both in physics and mathematics, a unified treatment of which may be found in [Collins et al. $\geq$ 2014]. We delay dealing with these averages for the moment and press on in our calculation.

We return to the expression

$$\begin{aligned}\tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)} \dots U_N X_N^{p(n)} U_N^* Y_N^{q(n)}] &= \frac{1}{N}\sum_{i,j,i',j'} \mathbb{E}\biggl[\prod_{k=1}^{n} U_N(i(k)j(k))\bar{U}_N(i'(k)j'(k))\biggr] \\ &\quad \times \prod_{k=1}^{n} X_N^{p(k)}(j(k)j'(k))Y_N^{q(k)}(i'(k)i\gamma(k)),\end{aligned}$$

and apply the Weingarten formula. The calculation evolves as follows:

$$\begin{aligned}&\tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)} \dots U_N X_N^{p(n)} U_N^* Y_N^{q(n)}] \\ &= \frac{1}{N}\sum_{i,j,i',j'}\sum_{\rho,\sigma\in\mathbf{S}(n)} \delta_{i\sigma,i'}\delta_{j\rho,j'}\mathbb{E}\biggl[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\rho^{-1}\sigma(k))\biggr] \\ &\qquad \times \prod_{k=1}^{n} X_N^{p(k)}(j(k)j'(k))Y_N^{q(k)}(i'(k)i\gamma(k)) \\ &= \frac{1}{N}\sum_{\rho,\sigma\in\mathbf{S}(n)} \mathbb{E}\biggl[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\rho^{-1}\sigma(k))\biggr] \\ &\qquad \times \sum_{i',j}\prod_{k=1}^{n} X_N^{p(k)}(j(k)j\rho(k))Y_N^{q(k)}(i'(k)i\sigma^{-1}\gamma(k))\end{aligned}$$

$$\begin{aligned}&= \frac{1}{N}\sum_{\rho,\sigma\in\mathbf{S}(n)} \mathbb{E}\bigg[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\rho^{-1}\sigma(k))\bigg] \\ &\qquad\qquad \times \operatorname{Tr}_{\rho}(X_N^{p(1)},\dots,X_N^{p(n)})\operatorname{Tr}_{\sigma^{-1}\gamma}(Y_N^{p(1)},\dots,Y_N^{p(n)}) \\ &= \sum_{\rho,\sigma\in\mathbf{S}(n)} \mathbb{E}\bigg[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\rho^{-1}\sigma(k))\bigg] N^{c(\rho)+c(\sigma^{-1}\gamma)-1}\operatorname{tr}_{\rho}(X_N^{p(1)},\dots,X_N^{p(n)}) \\ &\qquad\qquad \times \operatorname{tr}_{\sigma^{-1}\gamma}(Y_N^{p(1)},\dots,Y_N^{p(n)}).\end{aligned}$$

At this point we are forced to deal with the permutation correlators

$$\mathbb{E}\bigg[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\pi(k))\bigg].$$

Perhaps the most appealing presentation of these expectations is as a power series in N^{-1}. It may be shown [Novak 2010] that

$$\mathbb{E}\bigg[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\pi(k))\bigg] = \frac{1}{N^n}\sum_{r=0}^{\infty}(-1)^r\frac{c_{n,r}(\pi)}{N^r},$$

for any $\pi\in\mathbf{S}(n)$, where the coefficient $c_{n,r}(\pi)$ equals the number of factorizations

$$\pi = (s_1\ t_1)\dots(s_r\ t_r)$$

of π into r transpositions $(s_i\ t_i)\in\mathbf{S}(n),\ s_i<t_i$, which have the property that

$$t_1\le\cdots\le t_r.$$

This series is absolutely convergent for $N\ge n$, but divergent for $N<n$. This will not trouble us since we are looking for $N\to\infty$ asymptotics with n fixed. Indeed, let $|\pi| = n-c(\pi)$ denote the distance from the identity permutation to π in the Cayley graph of $\mathbf{S}(n)$. Then, since any permutation is either even or odd, we have

$$\begin{aligned}\mathbb{E}\bigg[\prod_{k=1}^{n} U_N(kk)\bar{U}_N(k\pi(k))\bigg] &= \frac{1}{N^n}\sum_{r=0}^{\infty}(-1)^r\frac{c_{n,r}(\pi)}{N^r} \\ &= \frac{(-1)^{|\pi|}}{N^{n+|\pi|}}\sum_{g=0}^{\infty}\frac{c_{n,|\pi|+2g}(\pi)}{N^{2g}} \\ &= \frac{a(\pi)}{N^{n+|\pi|}} + O\bigg(\frac{1}{N^{n+|\pi|+2}}\bigg),\end{aligned}$$

where $a(\pi) = (-1)^{|\pi|}c_{n,|\pi|}(\pi)$ is the leading asymptotics. We may now continue our calculation:

$$\begin{aligned}
&\tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)} \dots \\
&\qquad\qquad \tau_N[U_N X_N^{p(1)} U_N^* Y_N^{q(1)} \dots U_N X_N^{p(n)} U_N^* Y_N^{q(n)}] \\
&= \sum_{\rho,\sigma\in\mathbf{S}(n)} \left(\frac{a(\rho^{-1}\sigma))}{N^{n+|\rho^{-1}\sigma|}} + O\left(\frac{1}{N^{n+|\rho^{-1}\sigma|+2}}\right)\right) N^{c(\rho)+c(\sigma^{-1}\gamma)-1} \\
&\qquad\qquad \times \operatorname{tr}_\rho(X_N^{p(1)}, \dots, X_N^{p(n)}) \operatorname{tr}_{\sigma^{-1}\gamma}(Y_N^{p(1)}, \dots, Y_N^{p(n)}) \\
&\quad = \sum_{\rho,\sigma\in\mathbf{S}(n)} \left(a(\rho^{-1}\sigma) + O\left(\frac{1}{N^2}\right)\right) N^{|\gamma|-|\rho|-|\rho^{-1}\sigma|-|\sigma^{-1}\gamma|} \\
&\qquad\qquad \times \operatorname{tr}_\rho(X_N^{p(1)}, \dots, X_N^{p(n)}) \operatorname{tr}_{\sigma^{-1}\gamma}(Y_N^{p(1)}, \dots, Y_N^{p(n)}).
\end{aligned}$$

Putting everything together, we have shown that

$$\begin{aligned}
&\tau_N[U_N f_1(X_N) U_N^* g_1(Y_N) \dots U_N f_n(X_N) U_N^* g_n(Y_N)] \\
&\qquad = \sum_{\rho,\sigma\in\mathbf{S}(n)} \left(a(\rho^{-1}\sigma) + O\left(\frac{1}{N^2}\right)\right) N^{|\gamma|-|\rho|-|\rho^{-1}\sigma|-|\sigma^{-1}\gamma|} \\
&\qquad\qquad \times \operatorname{tr}_\rho(f_1(X_N), \dots, f_n(X_N)) \operatorname{tr}_{\sigma^{-1}\gamma}(g_1(Y_N), \dots, g_n(Y_N)),
\end{aligned}$$

and it remains to show that the $N \to \infty$ limit of this complicated expression is zero. To this end, consider the order $|\gamma| - |\rho| - |\rho^{-1}\sigma| - |\sigma^{-1}\gamma|$ of the ρ, σ term in this sum. The positive part, $|\gamma| = n - 1$, is simply the length of any geodesic joining the identity permutation to γ in the Cayley graph of $\mathbf{S}(n)$. The negative part, $-|\rho| - |\rho^{-1}\sigma| - |\sigma^{-1}\gamma|$, is the length of a walk from the identity to γ made up of three legs: a geodesic from id to ρ, followed by a geodesic from ρ to σ, followed by a geodesic from σ to γ. Thus the order of the ρ, σ term is at most N^0, and this occurs precisely when ρ and σ lie on a geodesic from id to γ; see Figure 16. Thus

$$\begin{aligned}
&\lim_{N\to\infty} \tau_N[U_N f_1(X_N) U_N^* g_1(Y_N) \dots U_N f_n(X_N) U_N^* g_n(Y_N)] \\
&\quad = \sum_{|\rho|+|\rho^{-1}\sigma|+|\sigma^{-1}\gamma|=|\gamma|} a(\rho^{-1}\sigma)\tau_\rho(f_1(X), \dots, f_n(X))\tau_{\sigma^{-1}\gamma}(g_1(Y), \dots, g_n(Y)).
\end{aligned}$$

Since

$$\tau[f_1(X)] = \tau[g_1(Y)] = \dots = \tau[f_n(X)] = \tau[g_n(Y)] = 0,$$

in order to show that the sum on the right has all terms equal to zero it suffices to show that the condition $|\rho| + |\rho^{-1}\sigma| + |\sigma^{-1}\gamma| = |\gamma|$ forces either ρ or $\sigma^{-1}\gamma$ to have a fixed point. This is because τ_ρ and $\tau_{\sigma^{-1}\gamma}$ are products determined by the cycle structure of the indexing permutation. Since ρ, σ lie on a geodesic id $\to \gamma$, we have $|\rho| + |\sigma^{-1}\gamma| \le |\gamma| = n - 1$, so that one of ρ or $\sigma^{-1}\gamma$ is a product of at most $(n - 1)/2$ transpositions. In the extremal case, all of these transpositions

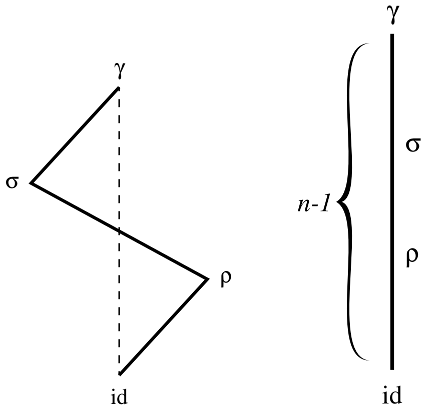

Figure 16. Only geodesic paths survive in the large N limit.

are joins, leading to a permutation consisting of an $(n-1)$-cycle and a fixed point.

3.7. ***GUE+GUE.*** Imagine that we had been enumeratively lazy in our construction of the GUE matrix model of a free semicircular pair, and had only shown that two iid GUE matrices $X_N^{(1)}, X_N^{(2)}$ are asymptotically free without determining their individual limiting distributions. We could then appeal to the free central limit theorem to obtain that the limit distribution of the random matrix

$$S_N = \frac{X_N^{(1)} + \cdots + X_N^{(n)}}{\sqrt{N}},$$

where the $X_N^{(i)}$ are iid GUE samples, is standard semicircular. On the other hand, since the matrix elements of the $X_N^{(i)}$ are independent Guassians whose variances add, we see that the rescaled sum S_N is itself an $N \times N$ GUE random matrix for each finite N. Thus we recover Wigner's semicircle law (for GUE matrices) from the free central limit theorem.

3.8. ***GUE+deterministic.*** Let X_N be an $N \times N$ GUE random matrix. Let Y_N be an $N \times N$ deterministic Hermitian matrix whose spectral measure ν_N converges weakly to a compactly supported probability measure ν. Let σ be the limit distribution of the random matrix $X_N + Y_N$. Since X_N, Y_N are asymptotically free, we have

$$\sigma = \mu \boxplus \nu,$$

where μ is the Wigner semicircle.

3.9. ***Randomly rotated + diagonal.*** Consider the $2N \times 2N$ diagonal matrix

$$D_{2N} = \begin{bmatrix} 1 & & & & \\ & -1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & -1 \end{bmatrix}$$

whose diagonal entries are the first $2N$ terms of an alternating sequence of ± 1, all other entries being zero. Let U_{2N} be a $2N \times 2N$ CUE random matrix, and consider the random Hermitian matrix

$$A_{2N} = U_{2N} D_{2N} U_{2N}^* + D_{2N}.$$

Let μ_{2N} denote the spectral measure of A_{2N}. We claim that μ_{2N} converges weakly to the arcsine distribution

$$\mu(\mathrm{d}t) = \frac{1}{\pi\sqrt{4-t^2}}\,\mathrm{d}t, \qquad t \in [-2, 2],$$

as $N \to \infty$.

Proof: Set $X_{2N} = U_{2N} D_{2N} U_{2N}^*$ and $Y_{2N} = D_{2N}$. Then X_N, Y_N is a random matrix model for a pair of free random variables X, Y each of which has the ± 1-Bernoulli distribution

$$\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}.$$

Thus the limit distribution of their sum is

$$\text{Bernoulli} \boxplus \text{Bernoulli} = \text{Arcsine}.$$

References

[Andrews et al. 1999] G. E. Andrews, R. Askey, and R. Roy, *Special functions*, Encyclopedia of Mathematics and its Applications **71**, Cambridge University Press, Cambridge, 1999.

[Bercovici and Voiculescu 1993] H. Bercovici and D. Voiculescu, "Free convolution of measures with unbounded support", *Indiana Univ. Math. J.* **42**:3 (1993), 733–773.

[Biane 1997] P. Biane, "On the free convolution with a semi-circular distribution", *Indiana Univ. Math. J.* **46**:3 (1997), 705–718.

[Biane 2002] P. Biane, "Free probability and combinatorics", pp. 765–774 in *Proceedings of the International Congress of Mathematicians, II* (Beijing, 2002), edited by T. Li, Higher Ed. Press, Beijing, 2002.

[Brézin et al. 1978] E. Brézin, C. Itzykson, G. Parisi, and J. B. Zuber, "Planar diagrams", *Comm. Math. Phys.* **59**:1 (1978), 35–51.

[Collins et al. ≥ 2014] B. Collins, S. Matsumoto, and J. Novak, *An invitation to Weingarten calculus*, In preparation.

[Connes 1994] A. Connes, *Noncommutative geometry*, Academic Press, San Diego, CA, 1994.

[Diaconis 2003] P. Diaconis, "Patterns in eigenvalues: the 70th Josiah Willard Gibbs lecture", *Bull. Amer. Math. Soc.* (*N.S.*) **40**:2 (2003), 155–178.

[Erdős et al. 2011] L. Erdős, B. Schlein, and H.-T. Yau, "Universality of random matrices and local relaxation flow", *Invent. Math.* **185**:1 (2011), 75–119.

[Etingof 2003] P. Etingof, "Mathematical ideas and notions of quantum field theory", 2003, http://math.mit.edu/~etingof/lect.ps. lecture notes.

[Fisher and Wishart 1932] R. A. Fisher and J. Wishart, "The derivation of the pattern formulae of two–way partitions from those of simpler patterns", *Proc. London Math. Soc.* **S2-33**:1 (1932), 195.

[Flajolet and Sedgewick 2009] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge Univ. Press, 2009.

[Forrester 2010] P. J. Forrester, *Log-gases and random matrices*, London Mathematical Society Monographs Series **34**, Princeton University Press, Princeton, NJ, 2010.

[Graham et al. 1989] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*, Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1989.

[Haagerup and Thorbjørnsen 2005] U. Haagerup and S. Thorbjørnsen, "A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group", *Ann. of Math.* (2) **162**:2 (2005), 711–775.

[Hald 1981] A. Hald, "T. N. Thiele's contributions to statistics", *Internat. Statist. Rev.* **49**:1 (1981), 1–20 (one plate).

[Hald 2000] A. Hald, "The early history of cumulants and the Gram–Charlier series", *International Statistical Review* **68**:2 (2000), 137–153.

[Hiai and Petz 2000] F. Hiai and D. Petz, *The semicircle law, free random variables, and entropy*, Amer. Math. Soc., Providence, RI, 2000.

[Hurwitz 1891] A. Hurwitz, "Über Riemann'sche Flächen mit gegebenen Verzweigungspunkten", *Mathematische Annalen* **39** (1891), 1–66.

[Kenyon and Okounkov 2007] R. Kenyon and A. Okounkov, "Limit shapes and the complex Burgers equation", *Acta Math.* **199**:2 (2007), 263–302.

[Kesten 1959] H. Kesten, "Symmetric random walks on groups", *Trans. Amer. Math. Soc.* **92** (1959), 336–354.

[Matytsin 1994] A. Matytsin, "On the large-N limit of the Itzykson–Zuber integral", *Nuclear Phys. B* **411**:2-3 (1994), 805–820.

[Mazur 2006] B. Mazur, "Controlling our errors", *Nature* **443** (2006), 38–39.

[Mingo and Speicher $\geq$ 2014] J. A. Mingo and R. Speicher, *Free probability and random matrices*, Fields Institute Monographs, Amer. Math. Soc., Providence, RI. To appear.

[Murty and Murty 2009] M. R. Murty and V. K. Murty, "The Sato–Tate conjecture and generalizations", pp. 639–646 in *Current trends in science: platinum jubilee special*, edited by N. Mukunda, Indian Academy of Sciences, Bangalore, 2009.

[Nestruev 2003] J. Nestruev, *Smooth manifolds and observables*, Graduate Texts in Mathematics **220**, Springer, New York, 2003.

[Nica and Speicher 2006] A. Nica and R. Speicher, *Lectures on the combinatorics of free probability*, London Mathematical Society Lecture Note Series **335**, Cambridge University Press, Cambridge, 2006.

[Novak 2010] J. I. Novak, “Jucys–Murphy elements and the unitary Weingarten function”, pp. 231–235 in *Noncommutative harmonic analysis with applications to probability, II*, edited by M. Bożejko et al., Banach Center Publ. **89**, Polish Acad. Sci. Inst. Math., Warsaw, 2010.

[Novak and Śniady 2011] J. Novak and P. Śniady, “What is... a free cumulant?”, *Notices Amer. Math. Soc.* **58**:2 (2011), 300–301.

[Pólya 1921] G. Pólya, “Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz”, *Math. Ann.* **84**:1-2 (1921), 149–160.

[Rota 1964] G.-C. Rota, “On the foundations of combinatorial theory, I: theory of Möbius functions”, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **2** (1964), 340–368.

[Roth 2009] M. Roth, “Counting covers of an elliptic curve”, 2009, http://www.mast.queensu.ca/~mikeroth/notes/covers.pdf.

[Saloff-Coste 2001] L. Saloff-Coste, “Probability on groups: random walks and invariant diffusions”, *Notices Amer. Math. Soc.* **48**:9 (2001), 968–977.

[Samuel 1980] S. Samuel, “U(N) integrals, $1/N$, and the De Wit–’t Hooft anomalies”, *J. Math. Phys.* **21**:12 (1980), 2695–2703.

[Shlyakhtenko 2005] D. Shlyakhtenko, “Notes on free probability theory”, preprint, 2005. arXiv 0504063

[Soshnikov 1999] A. Soshnikov, “Universality at the edge of the spectrum in Wigner random matrices”, *Comm. Math. Phys.* **207**:3 (1999), 697–733.

[Speed 1983] T. P. Speed, “Cumulants and partition lattices”, *Austral. J. Statist.* **25**:2 (1983), 378–388.

[Stanley 1999] R. P. Stanley, *Enumerative combinatorics*, vol. 2, Cambridge Studies in Advanced Mathematics **62**, Cambridge University Press, Cambridge, 1999.

[Stanley 2007] R. P. Stanley, “Increasing and decreasing subsequences and their variants”, pp. 545–579 in *International Congress of Mathematicians, I*, edited by J. L. V. Marta Sanz-Solé, Javier Soria and J. Verdera, Eur. Math. Soc., Zürich, 2007.

[Stanley 2013] R. P. Stanley, “Catalan addendum”, 2013, http://www-math.mit.edu/~rstan/ec/catadd.pdf.

[Tao 2010] T. Tao, “254A, notes 5: free probability”, 2010, http://terrytao.wordpress.com/2010/02/10/245a-notes-5-free-probability.

[Tao and Vu 2011] T. Tao and V. Vu, “Random matrices: universality of local eigenvalue statistics”, *Acta Math.* **206**:1 (2011), 127–204.

[Voiculescu 1986] D. Voiculescu, “Addition of certain noncommuting random variables”, *J. Funct. Anal.* **66**:3 (1986), 323–346.

[Voiculescu et al. 1992] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free random variables*, CRM Monograph Series **1**, American Mathematical Society, Providence, RI, 1992.

[Wigner 1958] E. P. Wigner, “On the distribution of the roots of certain symmetric matrices”, *Ann. of Math.* (2) **67** (1958), 325–327.

[Zvonkin 1997] A. Zvonkin, “Matrix integrals and map enumeration: an accessible introduction”, *Math. Comput. Modelling* **26**:8-10 (1997), 281–304.

jnovak@math.mit.edu

*Department of Mathematics, Massachusetts Institute
of Technology, 77 Massachusetts Avenue,
Cambridge, 02139-4307, United States*