MATH 233C: Topics in Combinatorics

Lecturer: Professor Persi Diaconis

Notes by: Andrew Lin

Spring 2025

1 March 31, 2025

This is a bit of an unusual course in that it's a course in combinatorics and probability and group theory, all shuffled together. The title of the course is "**enumeration, symmetry, and randomness**," and the main ideas are not so hard to say in a big-picture way. (Today's lecture will be an introductory one; precise definitions and results will come later.)

The big idea is that we're given some set of combinatorial objects and want to say something about it:

Example 1

It's a well-known fact that there are n^{n-2} labeled trees (often called **Cayley trees**) on n vertices, say rooted at vertex 1. For example for $n = 4$, there are 6 trees that are paths  straight down from the vertex, 6 trees of the shape  , 3 trees of the shape  , and 1 tree of the shape  . We can then ask questions like "how many leaves are there in a typical tree" (which is harder than we might think), "what is the typical depth or width," or "what is the degree distribution of a typical vertex." We'll be able to answer some of these and also explain "why people care."

That's the "enumeration" part of the course; we also might care about "enumeration under symmetry:"

Example 2

Take the notation in the example above. The permutations $S_{n-1} = \{\sigma \in S_n : \sigma(1) = 1\}$ act on these Cayley trees, and the orbits are called **Polya trees** (they are basically the unlabeled trees, or in other words the number of diagrams we drew in the above example). But there's no formula for the number of Polya trees, and answering those questions from above is harder.

More generally, we can let $\mathfrak{X}$ be any finite set and G any finite group acting on $\mathfrak{X}$. This group action splits $\mathfrak{X}$ into orbits $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \cdots \cup \mathcal{O}_k$, and we might ask questions like "how many orbits are there?" or "what is the typical size of an orbit?" or "is there an intelligent way to describe or name the orbits?" or (to help answer the earlier questions) "how can we pick an orbit at random?".

If we look up Polya theory (and find the classical Polya's theorem), we end up asking very similar questions. A typical question in that setting is as follows:

Example 3

We have 10 dice. How many ways are there to paint them red, white, and blue up to symmetry (permuting the dice and choosing which side shows up on each die)?

Elementary combinatorics books will phrase questions in this way, and we'll mostly talk in that language as well. The content of the course will involve developing **techniques for building generating functions** $\sum_{n=0}^{\infty} |\mathcal{C}_n| \frac{x^n}{n!}$ (this is the topic of **species**) and also **techniques for getting information from them** (such as the machinery of the **Boltzmann sampler**). But sometimes this is difficult to do algorithmically, and we'll discuss the computer science complexity of such questions as well.

Here's a general outline of what we'll be covering this quarter (each of these is a lecture or a few lectures):

1. Introduction and course outline (this lecture),
2. Careful discussion of groups acting on sets (orbit-stabilizer, Burnside's lemma, proof of Sylow's theorem),
3. Basic Polya theory (cycle indices, wreath products, de Bruijn's theorem),
4. Permutation enumeration as a look into probabilistic combinatorics (for example, if $\sigma \in S_n$ is written as a product of disjoint cycles and $a_i(\sigma)$ is the number of i-cycles, then for a uniformly random σ we can ask about the length of the longest cycle, number of fixed points, number of cycles, order, and so on, and we can prove limit theorems about the limiting distributions as $n \to \infty$),
5. The Burnside process, which is a strategy for choosing an orbit of a group action at random (a Markov chain in which starting at $x \in \mathfrak{X}$, we choose a uniform s fixing x, and then choose a uniform y that s fixes – it turns out the stationary distribution is inversely proportional to the size of the orbit) – in particular, we need to describe how to actually carry out this process in explicit examples.
6. Some background theory on Markov chain convergence,
7. The special case of the Burnside process with $\mathfrak{X} = C_2^n$ and $G = S_n$ (where the group acts by permuting coordinates and the orbits are the "levels" of how many ones there are) and obtaining sharp rates of convergence,
8. Background theory on coupling and applications to the case $\mathfrak{X} = C_k^n$ and $G = S_n$,
9. Some Fourier analysis on the hypercube and how it plays out in our problems,
10. The Burnside process of double cosets (considering the orbits for subgroups H, K of a group G and $H \times K$ acting on G via $s^{(h,k)} = h^{-1}sk$). For example, let $G = GL_n(\mathbb{F}_q)$ be the space of invertible $n \times n$ matrices in a finite field, and let $H = K$ be the Borel subgroup (invertible upper triangular matrices); the Bruhat decomposition (which is basically Gaussian elimination) says that GL_n is the disjoint union of BwB over permutation matrices w, so we get a "nice labeling" of the double cosets by permutations. There's also an analog of the Bruhat decomposition for not-necessarily-invertible matrices: we have $\mathrm{Mat}_n(\mathbb{F}_q) = \sqcup_{\hat{w}} B\hat{w}B$ with $\hat{w}$ contained in the **rook monoid**.
11. Contingency tables (an important special case used in statistics): let $G = S_n$ and $H = S_\lambda$ for some partition $\lambda = (\lambda_1, \cdots, \lambda_k)$ of n, where S_λ can only permute the first λ_1 things, the next λ_2 things, and so on. Similarly letting $K = S_\mu$, the double cosets $S_\lambda \backslash S_n / S_\mu$ are in bijection with contingency tables (which are matrices with fixed row and column sums),
12. Conjugacy classes via a group G acting on itself by conjugation; for example for $G = S_n$ these are indexed by partitions of n. We'll do enumeration of various permutation statistics here and relate this to how to generate partitions of large size (and does so quickly).

13. More abstractly, species theory (one of the "schools" of how to build and write down generating functions) and the associated category theory formalism. For example, letting the generating function for Polya trees be $P(x) = \sum_{n=1}^{\infty} |P_n| x^n$, we have
$$x e^{P(x)+\frac{1}{2}P(x^2)+\frac{1}{3}P(x^3)+\cdots},$$
and there's an automated machine for proving such results, and there's also techniques for extracting coefficients from such facts.

14. The Boltzmann sampler (taking a functional equation or differential equation and using it to draw random objects from P_n),

15. Frobenius groups.

Going back to logistics, the main assignment of the quarter is a single course project, in which we pick from a selected list of papers and write a paper about it. (These will usually use similar techniques or propose open questions related to what we do in lecture!)

2 April 2, 2025

We'll start the actual "lecturing" today. There's no official book for this course, but some references that might be useful are Suzuki's "Group Theory I" and Kerber's "Applied Finite Group Actions" (for the group theory), as well as any standard book on combinatorics (for Polya theory).

Definition 4

Let $\mathfrak{X}$ be a finite set and G a finite group. We say that G **acts on** $\mathfrak{X}$ if there is a mapping $\mathfrak{X} \times G \to \mathfrak{X}$, denoted $(x, s) \mapsto x^s$, such that for any $x \in \mathfrak{X}$ and any $s_1, s_2 \in G$, we have $(x^{s_1})^{s_2} = x^{s_1 s_2}$, and (letting e denote the identity of the group) $x^e = x$.

Of course, this definition also works for infinite groups and sets (for example the orthogonal group on $\mathbb{R}^n$), but we won't need that much here. Given any group action, we can define an equivalence relation

$$x \sim y \iff x^s = y \text{ for some } s \in G.$$

This equivalence relation splits $\mathfrak{X}$ into equivalence classes, which we call **orbits**, and we can write

$$\mathfrak{X} = \mathcal{O}_1 \cup \cdots \cup \mathcal{O}_k.$$

We'll use the abuse of notation $\mathcal{O}_x$ to denote the orbit that contains x.

As mentioned last time, we are often curious about the number of orbits under a group action, whether they have nice names, or whether they fit together into some moduli space. For example, we might ask whether there's nice notions of distances between two orbits.

Definition 5

A group G acts **transitively** on a set $\mathfrak{X}$ if there is only one orbit.

Definition 6

Let $\mathfrak{X}, \mathfrak{Y}$ be G-sets (that is, sets with G-actions). A **G-map** is a map $\lambda : \mathfrak{X} \to \mathfrak{Y}$ such that $\lambda(x^s) = (\lambda(x))^s$ for all $x \in \mathfrak{X}$ and $s \in G$. If λ is a bijection, we say that $\mathfrak{X}$ and $\mathfrak{Y}$ are **isomorphic** G-sets.

The first fundamental result in this subject is the following:

Theorem 7 (Orbit-stabilizer theorem)

Suppose G acts transitively on $\mathfrak{X}$. Then $\mathfrak{X}$ is isomorphic to G/U as G-sets, where $U = \{s \in G : x^s = x\}$ for some fixed x. (We often denote this set by G_x.) In other words, $\mathfrak{X}$ is isomorphic to the set of cosets $\{Ut\}$, where $(Ut)^s = Ust$. In particular,

$$|\mathfrak{X}| = \frac{|G|}{|U|} = [G : U].$$

Furthermore, if G/U_1 and G/U_2 are isomorphic for some subgroups U_1, U_2, then $U_1 = U_2^s = s^{-1}U_2 s$ for some s.

Proof. Fix some $x \in \mathfrak{X}$ and define U as above. The group splits into disjoint cosets as $G = \sqcup Ut_i$; define the map $\lambda : G/U \to \mathfrak{X}$ by

$$\lambda(Ut) = x^t.$$

We can check that this map is one-to-one and onto (surjective because the group acts transitively) and that it is a G-map. □

Example 8

Suppose $\mathfrak{X} = G$ and the group action is conjugation (meaning that $t^s = s^{-1}ts$ for any $s, t \in G$). The orbits of this group action are conjugacy classes, and the orbit-stabilizer theorem (applied to each orbit) can be restated as the **class equation**:

$$|G| = \sum_{i=1}^{k} [G : C_G(x_i)],$$

where x_i lies in the ith conjugacy class and $C_G(x) = \{s \in G : xs = sx\}$.

For example for $G = S_n$, the conjugacy classes are indexed by cycle types, or equivalently partitions of n (for example, the permutation $(1, 3, 2, 4, 5)$ in two-line notation can be rewritten as the product $(132)(4)(5)$ of disjoint cycles, corresponding to the partition $3 + 1 + 1$ of 5). We write that a permutation σ has a_i i-cycles (so that it has a_1 fixed points, a_2 transpositions, and so on, and $\sum i a_i = n$), or in shorthand we write that $\sigma = \prod_{i=1}^{n} i^{a_i(\sigma)}$.

What's nice is that permutations in cycle notation are easy to conjugate, and indeed we can check that conjugation preserves the shapes of the cycles (and that we can get from any permutation of a certain cycle type to any other one):

$$\pi = (i_1 \cdots i_a)(j_1 \cdots j_b)(k_1 \cdots k_c) \implies \pi^\sigma = (\sigma(i_1) \cdots \sigma(i_a))(\sigma(j_1) \cdots \sigma(j_b))(\sigma(k_1) \cdots \sigma(k_c)).$$

Theorem 9 (Cauchy)

Let $\lambda = \prod i^{a_i}$ be a partition of n. Then

$$|\text{conjugacy class } \lambda \text{ in } S_n| = \frac{n!}{Z_\lambda}, \quad Z_\lambda = \prod_{i=1}^{n} i^{a_i} a_i!.$$

Proof. Use the orbit-stabilizer theorem; the group acts transitively on this conjugacy class containing λ, and so the size of the class is $n!$ divided by the size of $(S_n)_x$. But in the λth conjugacy class, the conjugations that fix a given permutation with a_i i-cycles are those that rearrange the cycles (yielding the factor of $a_i!$) and cycle the labels within each i-cycle (yielding a factor of i per cycle). Multiplying all such factors yields the Z_λ above. □

Theorem 10 (Not Burnside lemma)
Let G act on $\mathfrak{X}$, and let $F(s) = |x : x^s = x|$ be the number of fixed points of the set under $s \in G$. Then the average number of fixed points satisfies

$$\mathbb{E}[F(s)] = \frac{1}{|G|}\sum_{s\in G} F(s) = \text{number of orbits of } \mathfrak{X} \text{ under } G.$$

Burnside was one of the first big group theorists, and this lemma is often attributed to him (and thus we will call it Burnside's lemma) even though it wasn't actually in his book.

Proof. We will count the quantity $M = |(x, s) \in \mathfrak{X} \times G : x^s = x|$ in two different ways. On the one hand, this is equal to $\sum_{s\in G} F(s)$ by summing first over the possible group elements. But we also have $M = \sum_x |G_x|$ (where G_x is the set of group elements that fix x), and we can write this as a sum over orbits

$$\sum_x |G_x| = \sum_{i=1}^{k}\sum_{x\in\mathcal{O}_i} |G_x| = \sum_{i=1}^{k}\frac{|G|}{|G_x|}|G_x| = k|G|,$$

where the blue equality uses the orbit-stabilizer theorem and that $|G_x|$ is constant on each $\mathcal{O}_i$. So $k|G| = \sum_s F(s)$, and rearranging yields the result. □

Example 11
One of the first theorems in probability is Montmort's theorem (from 1708), which calculates the expected number of fixed points of a uniform permutation $\sigma \in S_n$. The way Montmort stated it is by taking two decks of cards and turning over the top cards at once, counting the number of matches.

We can solve this by noting that S_n acts transitively on $\{1, \cdots, n\}$, so the expected number of fixed points is 1 by Burnside's lemma. But soon we'll discuss in this class the moments and the actual limiting distribution.

As an application of all of this, we'll prove the following:

Theorem 12 (Sylow)
Let p be a prime, and suppose we have a group G with $|G| = p^m n$, where $p \not| n$. Then (1) G has a subgroup of size p^m, which we call a **Sylow p-subgroup**, (2) all such subgroups are conjugate, (3) any p-group is contained in some Sylow p-subgroup, (4) the number of Sylow p-subgroups divides $|G|$ and is congruent to 1 mod p.

Note that nothing like this is true if p is not prime: for example, the alternating group A_5 has size 60 but there is no subgroup of size 15. What's interesting is that with only the size of the group, we get lots of natural subgroups (in fact p-groups); we can often go from these groups to information about the full group, and this captures much of modern group theory.

Example 13

Let $G = S_p$ for p prime. We have $|G| = p!$ and the highest power of p that divides this is p, so the cyclic group C_p is a Sylow p-subgroup. Any p-cycle generates a Sylow p-subgroup, and the number of such cycles is $(p-1)!$. **However**, note that different p-cycles can generate the same Sylow p-subgroup – in fact $(p-1)$ of them and thus there are $(p-2)!$ Sylow p-subgroups, which divides $|G| = p!$ and is indeed 1 mod p by Wilson's theorem.

Fact 14

Let G be any finite group. Suppose that for all p, the minimum number of generators of a Sylow p-subgroup is at most d. Then G itself is generated by at most $(d+1)$ elements.

This theorem is actually very hard to prove and requires the classification of finite simple groups, but the point is that we can state it very easily and we see that information about the Sylow p-subgroups tells us about the full group!

3 April 4, 2025

We'll start with a standard use of the decomposition of a G-set into orbits, proving the first part of Sylow's theorem:

Part of Wielandt's proof of Theorem 12. Let $\mathfrak{X}$ be the set of all size-p^m subsets of G, meaning that $|\mathfrak{X}| = \binom{|G|}{p^m}$. By properties of binomial coefficients, $|\mathfrak{X}|$ is relatively prime to p. On the other hand, $\mathfrak{X}$ splits into orbits under group action by right multiplication, and thus one of the orbits will have size not divisible by p.

Call that orbit $\mathcal{O}_i$. Then the orbit-stabilizer theorem says that $\mathcal{O}_i = G/S$ for some isotropy subgroup $S = S_A$, where A is a subset of size p^m. We know that S is a multiple of p^m (to cancel out the powers of p in $|G|$). On the other hand, pick any $a \in A$ (this is some group element). We have $|S| = |aS|$, but $|aS| \leq |A| = p^m$ (because by definition S sends a only to things in A). Thus we in fact have $|S| = p^m$ and we have found our Sylow p-subgroup.

There are group action proofs of the other parts as well – we can check Suzuki's book for the rest of the proofs. □

Example 15

Next, we'll compute the Sylow p-subgroups of S_n (since we'll use them later in the course).

First, we need the following useful computation:

Theorem 16

For any n and any prime p, we have p^m exactly dividing n if

$$m = \sum_{k=1}^{\infty} \left\lfloor \frac{n}{p^k} \right\rfloor = \frac{n - (a_0 + a_1 + \cdots + a_\ell)}{p},$$

where $a_0, a_1, \cdots, a_\ell$ are the digits of n in base p.

The first equality comes from counting the multiples of p, p^2, p^3, and so on. And the second equality comes from writing out the base-p representation of n and also $\lfloor \frac{n}{p^k} \rfloor$ and plugging everything in.

So this tells us the size of the Sylow p-subgroup we're looking for. We know that for $n = p$, we have $m = 1$ and the Sylow p-subgroup is C_p; similarly until $n = 2p - 1$ the Sylow is $C_p \times \text{id}$. But then when $n = 2p$, we want a Sylow p-subgroup of size $|S| = p^2$, which will be $C_p \times C_p$ (cycle among the first p numbers and also the last p).

That logic now holds up until $n = p^2 - 1$ (we have a product of cyclic groups together with some ientities), but then we get something interesting for $n = p^2$: we need $m = p+1$, and we get this from the subgroup $C_p^p \rtimes C_p = C_p \wr C_p$ (we can independently cyclically permute p different p-cycles, and then we can also cycle the order of those cycles) – these are called **chandelier groups**. Similarly for $n = p^j$ the Sylow p-subgroup is $C_p \wr C_p \wr \cdots \wr C_p$, so that we have a layer-j chandelier, and in general if $n = a_0 + a_1 p + a_2 p^2 + \cdots + a_\ell p^\ell$ we have a direct product of a_i copies of the layer-i chandelier.

Example 17

For a different kind of example, consider $G = GL_n(\mathbb{F}_q)$, the set of $n \times n$ matrices with entries in $\mathbb{F}_q$ (so $q = p^a$ is a prime power). The size of the group is the number of bases (thinking of this group acting on an n-dimensional space):

$$|G| = (q^n - 1)(q^n - q)(q^n - q^2) \cdots (q^n - q^{n-1}) = q^{\binom{n}{2}} \prod_{i=1}^{n} (q^i - 1).$$

since we must successively pick vectors that are linearly independent from the previous ones.

From the expression for $|G|$, we see that the Sylow p-subgroup must be of size $q^{\binom{n}{2}}$, and it's exactly the **uni-upper-triangular matrices** U with 1 on the diagonal, 0s below the diagonal, and arbitrary elements above.

We know the Sylow p-subgroups for many groups, and in various cases they will come up in the Markov chains that we will study. Moving closer to that now, we'll next consider **cycle indices**. Much of Polya theory falls into this language:

Definition 18

Let G be a subgroup of S_n (we can think of this as saying that G is thought to act on some finite set $\{1, \cdots, n\}$). For any $s \in G$ of this form, say that s has $a_i(s)$ i-cycles, and we write $s \sim \prod_{i=1}^{n} i^{a_i(s)}$. The **cycle index polynomial** is

$$Z_G(x_1, \cdots, x_n) = \frac{1}{|G|} \sum_{s \in G} \prod_{i=1}^{n} x_i^{a_i(s)} = \mathbb{E}_G \left[\prod_{i=1}^{n} x_i^{a_i(s)} \right].$$

In words, this is the generating function which encodes the sizes of cycles in a permutation.

Example 19

For the cyclic group $G = C_n \subseteq S_n$, we have

$$Z_G = \frac{1}{n} \sum_{d|n} \phi(d) x_d^{n/d},$$

where ϕ is the Euler totient function (that is, the number of integers at most d that are relatively prime to d).

For example if $n = 4$, the elements of C_4 in cycle notation are id, (1234), (13)(24), and (1432), so that

$$Z_{C_n}(x_1, x_2, x_3, x_4) = \frac{1}{4}(x_1^4 + x_2^2 + 2x_4).$$

Example 20

If we consider the whole permutation group $G = S_n$, then

$$Z_{S_n}(x_1, \cdots, x_n) = \frac{1}{n!} \sum_{\sigma \in S_n} \prod_{i=1}^{n} x_i^{a_i(\sigma)} = \frac{1}{n!} \sum_{\lambda \vdash n} \frac{1}{Z_\lambda} \prod_{i=1}^{n} x_i^{a_i(\lambda)}$$

where the partition λ has a_i parts of i and $Z_\lambda = \prod i^{a_i} a_i!$.

This is difficult to compute explicitly term-by-term, but there's a nice way of putting this together across different values of n:

Theorem 21 (Polya's theorem)

Define the formal power series in infinitely many variables

$$Z(t) = \sum_{n=0}^{\infty} t^n Z_{S_n}(x_1, \cdots, x_n).$$

Then $Z(t) = \exp\left(tx_1 + \frac{t^2}{2}x_2 + \frac{t^3}{3}x_3 + \cdots\right) = \prod_{i=1}^{\infty} \exp\left(\frac{t^i}{i}x_i\right)$.

Proof. We have the power series expansion $e^\lambda = \sum_{a=0}^{\infty} \frac{\lambda^a}{a!}$, so that

$$\prod_{i=1}^{\infty} \exp\left(\frac{t^i}{i}x_i\right) = \prod_{i=1}^{\infty} \left(\sum_{a_i=0}^{\infty} \frac{\left(\frac{t^i}{i}x_i\right)^{a_i}}{a_i!} \right)$$

(if we are concerned about convergence issues, we can set all of the x_is bigger than some large N to 1). If we then expand this out, we can collect powers of n in t. This yields

$$\sum_{n=0}^{\infty} t^n \sum_{\substack{a_0, \cdots, a_n \\ \sum i a_i = n}} \prod_i \frac{x_i^{a_i}}{i^{a_i} a_i!},$$

and now the inner sum is exactly the cycle index polynomial we just derived. □

In the next lecture, we'll show off a lot of the things we can do with this formula, and we'll just do a quick application now:

Example 22

Set $x_1 = x$ and $x_2 = x_3 = \cdots = 1$. Then

$$Z_{S_n}(x, 1, 1, \cdots) = \frac{1}{n!} \sum_{\sigma \in S_n} x^{a_1(\sigma)}$$

is the generating function for the number of fixed points of σ.

But the "grand" power series is

$$Z(t) = \exp\left(tx + \sum_{i=2}^{\infty} \frac{t^i}{i}\right) = \frac{e^{t(x-1)}}{1-t}$$

using the fact that $-\log(1-t) = \sum_{i\geq 1} \frac{t^i}{i}$. So if we differentiate Z_{S_n} in x and set $x = 1$, we find that

$$\frac{1}{n!}\sum_{\sigma\in S_n} a_1(\sigma) = 1$$

because this is just the average number of fixed points in a cycle. On the other hand, we can differentiate $Z(t)$ in x and set $x = 1$ and we get

$$\frac{t}{1-t} = t + t^2 + t^3 + \cdots,$$

and that agrees with what we just found (all of the t-coefficients are 1 for different ns). But we can get much more than that: if we differentiate k times before setting x equal to 1, for $Z_{S_n}(x)$ we end up with

$$\frac{1}{n!}\sum_{\sigma\in S_n} a_1(\sigma)(a_1(\sigma)-1)\cdots(a_1(\sigma)-k+1) = \mathbb{E}\left[(a_1(\sigma))_k\right]$$

(where $(a)_k$ denotes the falling factorial) and for $Z_t(x)$ we get $\frac{t^k}{1-t} = t^k + t^{k+1} + \cdots$. So equating sides again, this proves the following fact:

Theorem 23

Pick σ uniformly in S_n. Then

$$\mathbb{E}\left[(a_1(\sigma))_k\right] = \begin{cases} 1 & n \geq k \\ 0 & \text{otherwise.} \end{cases}$$

In particular, the variance of the number of fixed points is then also equal to 1, and in fact we get all of the moments (they're stable). That stability is not an accident, and it'll come up in some of the "category theory" later! And we can generally find this kind of formula not just for fixed points: we have

$$\mathbb{E}_{S_n}[(a_i)_k] = \begin{cases} \left(\frac{1}{i}\right)^k & n \geq ik \\ 0 & \text{otherwise.} \end{cases}$$

4 April 7, 2025

We'll spend today's lecture on some applications of Polya theory, seeing how it's used to solve interesting problems in probability. The main idea will be to start with the equation in Theorem 21 and plug in various interesting values for the variables t, x_i.

Definition 24

The **Poisson distribution** with parameter λ is the discrete random variable with probability mass

$$p_\lambda(j) = \frac{e^{-\lambda}\lambda^j}{j!}$$

for $j \in \mathbb{Z}_{\geq 0}$.

This indeed sums to 1 by the power series expansion of e^λ, and it turns out to be an interesting probability measure – for references, we can see Kingman's book "Poisson Processes" or Barbour, Holst, and Janson's "Poisson Approximation" (discussing the Chen-Stein method) for the material we discuss today. Here are some useful facts we'll need about this distribution:

- If $X \sim \text{Pois}(\lambda)$, then (letting x be a parameter) we get the formula for the generating function

$$\mathbb{E}\left[x^X\right] = \sum_{j=0}^{\infty} x^j \frac{e^{-\lambda}\lambda^j}{j!} = e^{-\lambda}e^{x\lambda} = e^{\lambda(x-1)}$$

  for any x.

- Differentiating this k times and then setting $x = 1$, we find that

$$\frac{\partial}{\partial x^k}\mathbb{E}\left[x^X\right]_{x=1} = \mathbb{E}\left[X(X-1)\cdots(X-k+1)\right] = \lambda^k;$$

  that is, the falling factorial moments of a Poisson distribution are nice.

- In particular, $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$ as well (so that's the real meaning of the parameter).

- If X, Y are independent Poisson variables with parameters λ, η respectively, then $X+Y$ is Poisson with parameter $\lambda + \eta$ (because the generating functions $e^{\lambda(x-1)}$ an $e^{\eta(x-1)}$ multiply to $e^{(\lambda+\eta)(x-1)}$).

Corollary 25

Multiply both sides of Polya's cycle index theorem by $1-t$, and on the right-hand side use that $1-t = e^{\log(1-t)} = e^{-t-\frac{t^2}{2}-\frac{t^3}{3}-\cdots}$. We then find that

$$\sum_{n=0}^{\infty}(1-t)t^n Z_{S_n}(x_1, \cdots, x_n) = \prod_{j=1}^{\infty}\exp\left(\frac{t^j}{j}(x_j - 1)\right).$$

The point is that this equation can be interpreted probabilistically! For the left-hand side, we can choose a "permutation of arbitrary length" with a specified distribution by first choosing N such that $\mathbb{P}(N = n) = (1-t)t^n$ (that is, flip a coin with probability t of being heads, and keep flipping until we get a tail – this is called the geometric distribution), and then pick $\sigma \in S_N$ uniformly at random. If the $\{A_i\}_{i=1}^{\infty}$s are the cycles of this (two-stage-sampled) σ, we know that $\sum iA_i = N$, but now N is random. The equality then says that $\{A_i\}_{i=1}^{\infty}$ are **exactly** independent, and each A_i is Poisson of parameter $\frac{t^i}{i}$. So **randomization makes the cycles independent**, and this is part of the topic of conditional limit theorems (this is somewhat similar to the **grand canonical ensemble** in statistical physics).

Thus if we care about some feature of permutation $\sigma \in S_n$ depending only on the cycle counts (conjugacy class), it is usually easy to understand that feature first for the $\{A_i\}$s, since they are independent. And as we'll see in some specific examples, sending $t \to 1$ sends $N \to \infty$, and what remains to show is that we also get the limit for S_n of a specific large n.

Remark 26. *Later on in the course, we'll discuss the Boltzmann sampler, which is a general way to sample a uniform object of a given n by doing this randomization procedure. In this case it's not so useful because we can easily generate uniform permutations, but we'll see later on cases where it is!*

Example 27

Consider the feature $a_i(\sigma)$, which is the number of i-cycles of the permutation σ.

In Polya's theorem above, set $x_i = x$ and all other variables x_j to 1. We then have (on the left-hand side) that the coefficient of t^n is

$$Z_{S_n}(1, \cdots, 1, x, 1, \cdots, 1) = \frac{1}{n!}\sum_{\sigma \in S_n} x^{a_i(\sigma)};$$

differentiating k times and then setting $x = 1$ yields $\mathbb{E}[a_i(a_i-1)\cdots(a_i-k+1)]$. And on the right-hand side, we have

$$\exp\left(\frac{t^i}{i}x + \sum_{j\neq i}\frac{t^j}{j}\right) = \frac{e^{\frac{t^i}{i}(x-1)}}{1-t}$$

using the same fact about $\log(1-t)$ as before; differentiating k times and then setting $x = 1$ for this yields

$$\frac{1}{1-t}\left(\frac{t^i}{i}\right)^k = \left(\frac{t^i}{i}\right)^k (1 + t + t^2 + \cdots).$$

So matching coefficients we indeed find the following (as promised at the end of last lecture):

Corollary 28

We have

$$\mathbb{E}_{S_n}[(a_i(\sigma))_k] = \begin{cases} \left(\frac{1}{i}\right)^k & n \geq ik \\ 0 & \text{otherwise.} \end{cases}$$

We can go back and forth between falling factorial moments and ordinary moments, and we thus find that

$$\mathbb{E}_{S_n}\left[a_i(\sigma)^k\right] = \mathbb{E}\left[X^k\right] \text{ for all } 1 \leq k \leq \frac{n}{i},$$

where X is Poisson of parameter $\frac{1}{i}$. And we can now make use of the language of classical limit theorems:

Definition 29

Let μ be any probability distribution on $\mathbb{R}$, and let $\mu_k = \int_{-\infty}^{\infty} x^k \mu(dx)$ be the kth moment of a random variable X distributed according to μ. We say that μ is **determined by its moments** if $\mu_k < \infty$ for all k, and if ν is another measure with moments $\nu_k = \mu_k$ for all k, then $\nu = \mu$.

In words, a measure is determined by its moments if it is the only probability measure with those moments.

Example 30

The normal distribution (of any mean and variance) is determined by its moments, and so is the Poisson distribution.

The idea is that if the generating function of μ is analytic in a neighborhood of 0, then it's determined, but there are other examples as well. For example if Y is standard normal and $X = Y^2$ (a chi-square distribution) then it is determined, but if $X = Y^3$ then it is not. (And exponent 4 is determined, while 5 and higher is not, and similarly a lognormal distribution e^Y is not.) In general, measures on compact sets are always determined.

Theorem 31 (Method of moments)

Suppose that μ is determined by its moments, and $\nu^{(n)}$ are probability measures on $\mathbb{R}$ whose kth moments $\nu_k^{(n)}$ converge to μ_k as $n \to \infty$ for any fixed k. Then $\nu^{(n)}$ **converges in distribution** to μ, meaning that $\mathbb{P}_{\nu^{(n)}}((-\infty, x]) \to \mathbb{P}_\mu((-\infty, x])$ for all x which are continuity points of μ, or equivalently $\mathbb{E}_{\nu^{(n)}}[f(X)] \to \mathbb{E}_\mu[f(X)]$ for any bounded continuous function f.

On the other hand, we might be curious what happens if the measure μ isn't determined. What goes wrong in that case is that for every $c \in \mathbb{R}$, there is some probability measure ν_c with

$$\int x^k \nu_c(dx) = \int x^k \mu(dx)$$

such that ν_c has an atom of positive mass. So in fact there is an uncountable family of measures – our measure is wildly non-unique in such cases. If we're curious to learn more about this material, we can see Professor Diaconis's "Application of the method of moments in probability and statistics." (This came out of a two-day conference on the "moment problem" – it really is big subject spanning different fields of math.)

One question we might then ask is "how we can tell" whether a measure μ is determined in general. There are sufficient conditions; for example (see Billingsley) if the generating function

$$f(Z) = \int x^Z \mu(dx)$$

is analytic in a complex neighborhood of 0, then the moments are the Taylor coefficients and knowing the Laplace transform gives us the measure. Additionally, the standard "fancy condition" is **Carleman's condition**, which says that it is sufficient to have

$$\sum_{k=1}^{\infty} \frac{1}{(\mu_{2k})^{1/2k}} = \infty.$$

(For example, this does indeed diverge for the standard normal, and it works for chi-square with 1 degree of freedom, and it fails for Z^3.) But there isn't a general classification known.

But the point is that if we go back to Poisson stuff, we showed today that if $\nu^{(n)}$ is the probability distribution for $a_i(\sigma)$ with $\sigma \in S_n$, then $\nu_k^{(n)} \to \mu_k$ for μ Poisson with parameter $\frac{1}{i}$ (in fact, the moments are eventually exactly equal). Thus

$$\mathbb{P}_{S_n}(a_i(\sigma) = j) \to \mathbb{P}\left(\text{Pois}\left(\frac{1}{i}\right) = j\right) = \frac{e^{-1/i}\left(\frac{1}{i}\right)^j}{j!}.$$

And in fact we can do the exact same argument (differentiating in an appropriate way but keeping multiple of the variables at once) to show the following:

Theorem 32

For every L and every $k_1, \cdots, k_L \in \mathbb{N}$, we have convergence of the joint mixed moments

$$\mathbb{E}_{S_n}\left[a_1^{k_1} a_2^{k_2} \cdots a_L^{k_L}\right] = \mathbb{E}\left[X_1^{k_1} \cdots X_L^{k_L}\right] = \prod_{i=1}^{L} \mathbb{E}\left[X_i^{k_i}\right],$$

where X_is are independent Poisson($\frac{1}{i}$), as long as $n \geq \sum_{i=1}^{L} i k_i$.

So it's a bit strange that moments converge and then are "equal forever" – we say that the moments stabilize, and stabilization is a big deal in some parts of modern topology and other areas. And if we'd like to see more about this, we can take a look at Church, Ellenberg, and Farb's "FI-modules and stability for representations of symmetric groups;" the fact about Poisson variables is at the heart of all of this!

5 April 9, 2025

We'll start today with another use of the Polya index theorem:

Example 33

Let $C(\sigma) = \sum_{i=1}^{n} a_i(\sigma)$ be the number of cycles in a permutation σ. Observe that if we set all $x_i = x$, then the product $\prod_{i=1}^{n} x_i^{a_i(\sigma)}$ becomes $x^{C(\sigma)}$, so that

$$Z_{S_n}(x, x, \cdots, x) = \frac{1}{n!} \sum_{\sigma \in S_n} x^{C(\sigma)}$$

is the generating function for the number of cycles.

We can thus compare this to the right-hand side – we get

$$\prod_{i=1}^{\infty} \exp\left(\frac{t^i}{i} x\right) = \frac{1}{(1-t)^x}$$

by the usual Taylor expansion of log, and now by Newton's binomial expansion this expression simplifies to

$$\sum_{j=0}^{\infty} \binom{-x}{j} (-t)^j = \sum_{j=0}^{\infty} \frac{x(x+1)\cdots(x+j-1)}{j!} t^j,$$

and thus we find the answer (canceling out the factors of $n!$)

$$\sum_{\sigma \in S_n} x^{C(\sigma)} = x(x+1)\cdots(x+n-1).$$

This formula is famous – the coefficients of this polynomial are the signless Stirling numbesr, and once we know it we can easily prove it by induction in other ways. Furthermore, if we divide back by $n!$, we find that

$$\frac{1}{n!} \sum_{\sigma \in S_n} x^{C(\sigma)} = x\left(\frac{1}{2} + \frac{x}{2}\right) + \left(\frac{2}{3} + \frac{x}{3}\right) \cdots \left(1 - \frac{1}{n} + \frac{x}{n}\right).$$

But now $\left(1 - \frac{1}{i} + \frac{x}{i}\right)$ is the generating function of a Bernoulli random variable with parameter $\frac{1}{i}$; that is, $X_i = 0$ with probability $1 - \frac{1}{i}$ and $X_i = 1$ with probability $\frac{1}{i}$. Thus, we get the following probabilistic interpretation:

Corollary 34

Pick σ uniformly in S_n. Then $C(\sigma)$ has the same distribution as $X_1 + \cdots + X_n$ with independent $X_i \sim \text{Ber}\left(\frac{1}{i}\right)$.

Linearity of expectation then tells us that

$$\mathbb{E}_{S_n}[C(\sigma)] = 1 + \frac{1}{2} + \cdots + \frac{1}{n} = \log n + \gamma + O\left(\frac{1}{n}\right)$$

and (variances add for indepenent random variables)

$$\begin{aligned} \text{Var}_{S_n}(C(\sigma)) &= \sum_{i=1}^{n} \frac{1}{i}\left(1 - \frac{1}{i}\right) \\ &= \sum_{i=1}^{n} \frac{1}{i} - \frac{1}{i^2} \\ &= \log n + \gamma - \frac{\pi^2}{6} + O\left(\frac{1}{n}\right). \end{aligned}$$

Furthermore, the central limit theorem holds (even though our variables are not identically distributed), so that

$$\mathbb{P}_{S_n}\left(\frac{C(\sigma)-\log n}{\sqrt{\log n}} \le x\right) \to \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-x^2/2}dx.$$

So a typical permutation on 52 cards has around 4 or 5 cycles with standard deviation around 2.

Remark 35. *We could have also gotten this last theorem from our formula directly. Specifically we can use the measure where we pick a uniform permutation on S_N with $\mathbb{P}(N=n) = (1-t)t^n$ and note that the number of i-cycles is Poisson with parameter $\frac{t^i}{i}$. (To be more precise, we're really sampling from the union of all S_ns, and any permutation comes with an associated n so that the identity permutations are not all the same.) We now have potentially infinitely many A_is, but*

$$\mathbb{P}(A_i > 0) = 1 - e^{-t^i/i} \approx \frac{t^i}{i}$$

has a finite sum and thus the number of cycles is finite almost surely by Borel-Cantelli. That means that $C(\sigma) = \sum_{i=1}^{\infty} A_i$ makes sense under this "grand" measure, and its mean is then

$$\mathbb{E}_t[C(\sigma)] = \sum_{i=1}^{\infty}\frac{t^i}{i} = -\log(1-t),$$

which indeed diverges as $t \to 1$. Similarly the variance of $C(\sigma)$ can be calculated to be $-\log(1-t)$, and if we look at the usual proof of the Lindeberg central limit theorem we indeed get

$$\mathbb{P}_t\left(\frac{C(\sigma)+\log(1-t)}{\sqrt{-\log(1-t)}} \le x\right) \to \Phi(x)$$

under the randomizing-n measure. But we want to go from random n to fixed n (we know about the behavior of a power series and want to know about its coefficients), and we can do that directly but it takes a bit of work (specifically Tauberian theorems) – for a reference, we can see Shepp and Lloyd's "Ordered Cycle Lengths in a Random Permutation."

Example 36

We'll instead do something that in a sense "only Professor Diaconis can do:" specifically, we'll answer the question of "who cares about all of this." One motivation is the study of metrics on permutations – given two sets of rankings, we might care about how similar they are and how to measure a distance.

Six examples are the Hamming distance

$$d_H(\sigma,\eta) = \#\{i : \sigma(i) \neq \eta(i)\},$$

Kendall's tau (probably the most common one in practice, and somewhat a Riemannian metric)

$$d_\tau(\sigma,\eta) = \text{minimum number of adjacent transpositions to bring } \sigma \text{ to } \eta,$$

the Cayley distance

$$d_C(\sigma,\eta) = \text{minimum number of (any) transpositions to bring } \sigma \text{ to } \eta,$$

Spearman's rho

$$d_\rho(\sigma,\eta) = \sqrt{\sum_{i=1}^{n}(\sigma(i)-\eta(i))^2},$$

Spearman's footrule (using the ℓ^1 norm instead of the ℓ^2)

$$\sum_{i=1}^{N} |\sigma(i) - \eta(i)|,$$

and the Ulam distance (Don Knuth's favorite)

$$\text{minimum number of insertion-deletion operations to bring } \sigma \text{ to } \eta,$$

where an insertion-deletion operation is basically a cycle (if we put a card in the 5th position into the 10th, then we get the cycle $(5, 10, 9, 8, 7, 6)$) – that is, we look at the length metric under the generating set $S = \{(i, i+1, \cdots, j)\}$.

Remark 37. *We know how to calculate Kendall's tau between two permutations σ, η efficiently where we are only allowed to multiply the transpositions on the left (or just on the right). But if we allow ourselves to multiply either on the left or on the right at the same time, it's not even clear how to calculate it (we're allowed to either switch two adjacent positions or two adjacent values).*

All six of these are metrics on S_n (meaning that $d(\sigma, \sigma) = 0$, $d(\sigma, \tau) = d(\tau, \sigma)$, and $d(\sigma, \tau) \leq d(\sigma, \eta) + d(\eta, \tau)$), but importantly some but not all have invariance properties. For example, we wouldn't want the distance of similarity between two rankings to depend on the order in which the elements are listed – here we think of σ as a mapping from names to $\{1, \cdots, n\}$, and we want **right-invariance** $d(\sigma, \tau) = d(\sigma\eta, \tau\eta)$ where η is a permutation on the names. Similarly, it can also make sense to have **left-invariance** $d(\eta\sigma, \tau\sigma) = d(\sigma, \tau)$ or **bi-invariance** $d(\eta\sigma\eta', \eta\tau\eta') = d(\sigma, \tau)$. Only Hamming distance d_H and Cayley distance d_C are bi-invariant (to make Kendall's tau right-invariant we need to compute the minimum number of transpositions from σ^{-1} into τ^{-1} instead), and trying to find other bi-invariant metrics is important.

The point is that people do use these metrics:

- If $\sigma_1, \cdots, \sigma_N$ are N different rankings of n things (for example if N people made a ranking of 5 flavors of chocolate chip cookies, or in an election we're ranking a slate of candidates), we might want an idea of a "typical" or "mean" ranking. One way to do that is to choose σ^* so that $\sum_{i=1}^{n} d(\sigma_i, \sigma^*)$ is minimal.
- In statistics, we often do two-sample tests; for example given two sets of $\mathbb{R}$-valued data $\{X_1, X_2, \cdots, X_n\}$ and $\{Y_1, Y_2, \cdots, Y_m\}$, we often care whether the distributions are different. One strategy is to pass to rankings – the **Mann-Whitney test** combines the data and ranks those $n + m$ numbers, and it asks "how many adjacent transpositions does it take to bring all Xs together" (Kendall's tau between the two rankings).
- In psychophysical experiments, often people are shown seven shades of red and asked to rank them in brightness. There's a large literature on the effects of outside influences, and evaluating these outcomes often uses things like Spearman's statistics.
- Finally, we may be interested in building non-uniform distributions on the permutations S_n (for example, in the case above we expect the ranking of those seven shades to be close to the right answer, possibly with a few switches). One such example for a distribution "peaked at a location parameter σ_*" is

  $$\mathbb{P}_\beta(\sigma) = Z^{-1} e^{-\beta d(\sigma, \sigma_*)}$$

  for a normalizing constant Z. Given data, we might then want to estimate β and σ_*. For more literature on this, we can read Professor Diaconis' book "Group representations in probability and statistics" or Marden's "Analyzing and Modeling Rank Data."

A natural question we might ask at this point is then the following: if σ, τ are picked from some distribution, say uniform, what is the distribution of $d(\sigma, \tau)$? This question can be easily answered in some cases under the uniform distribution (the two bi-invariant ones), since we can just take σ to be the identity. Specifically, for the Hamming distance

$$d_H(\mathrm{id}, \tau) = n - a_1(\tau)$$

and we know the distribution of a_1 (Poisson with parameter 1): $\mathbb{P}(n - d_H(\mathrm{id}, \tau)) \approx \frac{1}{ej!}$. And for the Cayley distance, it turns out $d_C(\mathrm{id}, \tau) = n - C(\tau)$, so by what we've discussed today this distance is normal.

The other metrics are then not only functions of the conjugacy class – we know each of them, but each is a separate little theorem. For example, the number of required insertion-deletions has mean $2\sqrt{n}$ and standard deviation $n^{1/3}$, and in fact the fluctuations follow a Tracy-Widom distribution (this is due to Baik, Deift, and Johansson, and it's a very hard result).

Fact 38

There are some interesting open problems in this direction. We already mentioned the problem about allowing adjacent transpositions on either side simultaneously (for example, trying to calculate the mean). For another question, the two bi-invariant metrics we described are both quite concentrated (at order constant and $\log n$ for Hamming and Cayley, respectively); it would be interesting to find a two-sided metric which is "most spread out" and still sensible. And finally, we can consider the permutation statistic

$$D(\sigma) = \text{number of descents of } \sigma,$$

where σ has a **descent** at i if $\sigma(i+1) < \sigma(i)$. The math of descents is as rich as that of cycles (and in fact we can say more about joint distributions), and the problem is to make a right-invariant metric out of D. (Doing something like $d(\sigma, \tau) = d(1, \tau\sigma^{-1})$ is not symmetric, and then if we try to symmetrize we fail the triangle inequality.)

6 April 11, 2025

We've been doing cycle indices, and last time we showed some examples where they're interesting and where they can be computed. We'll see much more of this today, and this often goes under the name "plethysm."

Definition 39

Fix integers k, n, and let $\Gamma \subseteq S_k$ and $H \subseteq S_n$ be subgroups. Let $G = \Gamma^n \rtimes H = \Gamma \mathrm{wr} H$ be the **wreath product**, which is a subgroup of S_{kn} with elements of the form

$$\sigma = (\gamma_1, \cdots, \gamma_n; h), \quad \gamma_i \in \Gamma, h \in H.$$

Elements of G act as permutations on $1, \cdots, kn$ by letting γ_1 act on the first k places, γ_2 on the next k, and so on, and then letting h permute the blocks.

For example, consider $C_2^3 \rtimes S_3$. This is a subgroup of S_6, and if $\sigma = ((12), (1)(2), (12); (312))$ (in cycle notation), we first switch 1 and 2 and also 5 and 6, and then we send block 3 go to block 1, block 1 to block 2, and block 2 to block 3. We thus end up with the permutation 652134 in one-line notation, and thus in cycle notation we end up with $(164)(253) \in S_6$.

There are some "famous wreath products" that we should know about:

Example 40

The **hyperoctahedral group** B_n can be written as $C_2^n \rtimes S_n$; this is the group of symmetries of the hypercube (since each C_2 corresponds to flipping in one coordinate, and S_n permutes the coordinates), and it's also the set of permutations in S_{2n} which have "central symmetry" (meaning that $\sigma(i)+\sigma(2n+1-i) = 2n+1$ for all i). This is one of the seven irreducible Weyl groups (of the orthogonal group) and so it's a finite Coxeter group. More concretely, we can think of the elements of B_n as the set of signed $n \times n$ permutation matrices (meaning that we have one nonzero entry per row and column, and it is either 1 or -1).

Example 41

A **generalized permutation group** takes the form $C_k^n \rtimes S_n$ (so within each segment we can only cycle). More concretely, this can be represented as the set of $n \times n$ matrices with one nonzero entry per column, and it has to be one of the k roots of unity.

One place where this comes up is that given a permutation $\sigma \in S_n$, the centralizer of S_n in σ (that is, the set of all τ such that $\sigma\tau = \tau\sigma$) is

$$C_{S_n}(\sigma) = \prod_{i=1}^{n} C_i^{a_i} \rtimes S_{a_i}, \quad \sigma \sim 1^{a_1} \cdots n^{a_n}.$$

In other words, we look at all of the cycles of a certain size, and we're allowed to permute them and cycle each of them. Then conjugation is equivalent to relabeling the points, so we need to preserve the property "being in an i-cycle." We'll need this because we'll soon want to study the commuting graph of S_n.

Example 42

Finally, we can consider the full wreath product $S_k^n \rtimes S_n$. This is a maximal proper subgroup of S_{kn} (this is due to the classification of subgroups given by the O'Nan-Scott theorem). It also comes up in **ANOVA (analysis of variance)** in statistics – for example, if there are n classes of k students each, and each class is given a different teacher or textbook and then each class is given an exam, we might ask whether those treatments made a difference. And one way of analyzing this uses symmetries of the data, and the symmetry group is exactly this group. (So then using representation theory, we can understand the classical normal theory analysis.)

In particular, our first example B_n is a special case of this, and also if we have (for example) ten dice with faces painted in various colors considered up to symmetry, then the symmetry group turns out to be $S_4^{10} \rtimes S_{10}$ (because the group of symmetries of a die is the permutation group of the space diagonals).

In general if $G = \Gamma^n \rtimes H$, the size of the group is given by $|G| = |\Gamma|^n |H|$ (so the size of the octahedral group is $2^n n!$, for example). We can then ask the usual questions of "what does a random such permutation look like" (for example, the cycle count or descent count or length of the longest increasing subsequence).

Theorem 43 (Polya's plethysm theorem)

Let $G = \Gamma^n \rtimes H$ for $\Gamma \subseteq S_k$ and $H \subseteq S_n$. Then the cycle index polynomial satisfies

$$Z_G(x_1, \cdots, x_{kn}) = \frac{1}{|G|} \sum_{\sigma \in G} \prod_{i=1}^{n} x_i^{a_i(\sigma)} = Z_H(t_1, \cdots, t_n),$$

where $t_i = Z_\Gamma(x_i, x_{2i}, x_{3i}, \cdots, x_{ki})$. (This operation is what's called the **plethysm** of these two generating functions.)

For example, $C_2^2 \rtimes S_2$ is a subgroup of S_4 with $2^2 \cdot 2! = 8$ elements, and in one-line notation those elements are

$$1234, 2134, 1243, 2143, 3412, 3421, 4312, 4321.$$

The cycle types of these elements are $1^4, 1^2 2, 1^2 2, 2^2, 2^2, 4, 4, 2^2$, so that

$$Z_G(x_1, x_2, x_3, x_4) = \frac{1}{8}(x_1^4 + 2x_1^2 x_2 + 3x_2^2 + 2x_4).$$

Meanwhile, $Z_{C_2}(x_1, x_2) = Z_{S_2}(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2)$, and this theorem says that

$$\begin{aligned} Z_G(x_1, x_2, x_3, x_4) &= \frac{1}{2}\left(t_1^2 + t_2\right) \\ &= \frac{1}{2}\left(\left[\frac{1}{2}(x_1^2 + x_2)\right]^2 + \frac{1}{2}(x_2^2 + x_4)\right) \\ &= \frac{x_1^4}{8} + \frac{2x_1^2 x_2}{8} + \frac{x_2^2}{8} + \frac{2x_2^2}{8} + \frac{2x_4}{8}, \end{aligned}$$

which indeed checks out.

Example 44

We'll now illustrate how randomness can be made out of this theorem. If we fix k, n and let $G = \Gamma^n \rtimes S_n$, then we can pick $\sigma \in G$ uniformly at random. We want to calculate what the cycle count looks like; recall that in the case where Γ is the identity, we have Goncharov's theorem which says that $C(\sigma)$ is roughly normal with mean and variance $\log n$.

Recall that for any permutation group G (say contained in S_N), we can set all variables equal and get

$$Z_G(x, x, \cdots, x) = \frac{1}{|G|} \sum_{\sigma \in G} x^{a_1 + \cdots + a_N} = \mathbb{E}_G[x^{C(\sigma)}],$$

the generating function for the number of cycles. So plugging in x into Polya's plethysm theorem, we get

$$Z_G(x, \cdots, x) = Z_H(t_1, \cdots, t_n), \quad t_i = Z_\Gamma(x, \cdots, x) = \mathbb{E}_\Gamma[x^{C(\sigma)}].$$

Therefore translating to probability, we find the following fact:

Theorem 45

We have

$$Z_G(x, \cdots, x) = \mathbb{E}_G[x^{C(\sigma)}] = \sum_{j=1}^{n} p_H(j) C_\Gamma(x)^j,$$

where $p_H(j) = \mathbb{P}_H(C(\sigma) = j)$ is the probability that a random permutation in H has j cycles, and $C_\Gamma(x) = \mathbb{E}_\Gamma[x^{C(\sigma)}]$.

Now since $C_\Gamma(x)^j$ is the product of generating functions, we can interpret it as a convolution (that is, sum of independent random variables):

Corollary 46

Pick $\sigma \in \Gamma^n \rtimes H$ uniformly at random. Then we have the equality in distribution

$$C \stackrel{d}{=} \sum_{i=1}^{N} X_i$$

where X_i are iid from the generating function $C_\Gamma(x)$ and N is independent of the X_is with generating function $C_H(x)$.

That is, the number of cycles has the distribution of what's called a **randomly stopped sum** (we take a random number of iid samples and add them together); this is a standard subject in sequential analysis (where we possibly want to stop trials early if we notice a large effect). And this formula is useful for us, because by **Anscombe's theorem** this means that this sum is approximately normal with the appropriate mean and variance.

7 April 14, 2025

Today, we'll be continuing to study the cycles in $G = \Gamma^n \rtimes H$ for Γ, H subgroups of S_k, S_n respectively (where H acts on permutations on the n copies of Γ). Last time, we showed that $C(\sigma) = \sum a_i(\sigma)$ can be extracted from the cycle index polynomial of G, which itself can be extracted from the generating functions of Γ and H (see Theorem 43 and Corollary 46 for the precise statements); the key fact is that C has the same law as a randomly stopped sum of iid random variables X_i sampled according to C_Γ, where the stopping time is independently sampled according to C_H.

We can thus calculate by Wald's theorem that

$$\mathbb{E}[C_G] = \mathbb{E}[N]\mathbb{E}[X_1],$$

and we can also use the identity $\mathrm{Var}(W) = \mathbb{E}[\mathrm{Var}(W|X)] + \mathrm{Var}(\mathbb{E}[W|X])$, conditioning the cycle count $(W = C)$ on the stopping time $(X = N)$ to find that

$$\mathrm{Var}(C_G) = \mathbb{E}_H[N]\mathrm{Var}_\Gamma(C) + \mathbb{E}_H[N]^2\mathrm{Var}_\Gamma(C),$$

since the first term uses that a sum of n iid copies of X has variance $n\mathrm{Var}(X)$, and the second uses that $\mathbb{E}[C|N] = N\mathbb{E}[X_1]$. But the point is that all terms on the right-hand side can be computed – for example if $H = S_N$, then we've already seen that $\mathbb{E}_H[C] = H_n = \log n + \gamma + O\left(\frac{1}{n}\right)$ and $\mathrm{Var}_H(C) = \log n + \gamma - \frac{\pi^2}{6} + O\left(\frac{1}{n}\right)$, and if $H = C_n$ we also get similar nice formulas.

Theorem 47

For $\Gamma \subseteq S_k$ (with k fixed) and $H = S_n$ (we'll take n growing), let $G = \Gamma^n \rtimes S_n$ and let σ be uniform on G. Then

$$\mathbb{P}_G\left(\frac{C(\sigma) - \mu_n}{\tilde{\sigma}\sqrt{\log n}} \leq x\right) \xrightarrow{n\to\infty} \Phi(x),$$

where $\mu_n = \tilde{\mu}(\log n + \gamma + O(\frac{1}{n}))$ and $\tilde{\mu}, \tilde{\sigma}^2$ are the mean and variance of $C(\eta)$ for η sampled from Γ.

Proof. Let $X_1, X_2, \cdots$ be iid samples from the generating function $C_\Gamma(x)$ (so pick random elements from Γ and consider the cycle counts), **with the mean subtracted off** so that X_i are all mean zero. Let N_n have the probability distribution coming from $C_{S_n}(x)$. Goncharov's theorem implies that $\frac{N_n}{\log n} \to 1$ in probability. If we let $n_0 = \lfloor \log n \rfloor$, then letting $S_m = X_1 + \cdots + X_m$ (so that $\frac{S_m}{\sqrt{m}}$ converges to $N(0, \sigma^2)$)

$$\frac{S_{N_n}}{\sqrt{N_n}} = \left(\frac{S_{n_0}}{\sqrt{n_0}} + \frac{S_{N_n} - S_{n_0}}{\sqrt{n_0}}\right)\left(\frac{n_0}{N_n}\right)^{1/2},$$

and it just suffices to prove that $\frac{S_{N_n} - S_{n_0}}{\sqrt{n_0}} \to 0$ in probability. Indeed, for any $\varepsilon \in (0, \frac{1}{2})$ and defining

$$n_1 = \lfloor n_0(1-\varepsilon^3) \rfloor + 1, \quad n_2 = \lfloor n_0(1+\varepsilon^3) \rfloor,$$

it suffices to show the vanishing of the probability

$$\begin{aligned}\mathbb{P}\left(|S_{N_n} - S_{n_0}| > \varepsilon n_0^{1/3}\right) &= \mathbb{P}\left(S_{N_n} - S_{n_0}| > \varepsilon n_0^{1/3} \text{ and } N_n \in [n_1, n_2]\right) + \mathbb{P}\left(S_{N_n} - S_{n_0}| > \varepsilon n_0^{1/3} \text{ and } N_n \notin [n_1, n_2]\right) \\ &\leq \mathbb{P}\left(\max_{n_1 \leq k \leq n_0} |S_k - S_{n_0}| > \varepsilon n_0^{1/3}\right) + \mathbb{P}\left(\max_{n_0 \leq k \leq n_2} |S_k - S_{n_0}| > \varepsilon n_0^{1/3}\right) + \mathbb{P}\left(N_n \notin [n_1, n_2]\right).\end{aligned}$$

Now Kolmogorov's inequality lets us bound the maximum of a bunch of independent random variables of mean zero: the first two terms are bounded by $\frac{(n_0 - n_1)\tilde{\sigma}^2}{n_0}$ and $\frac{(n_2 - n_0)\tilde{\sigma}^2}{n_0}$, respectively, and by definition those last two terms are at most $2\varepsilon\sigma^2$. Thus we have

$$\mathbb{P}\left(|S_{N_n} - S_{n_0}| > \varepsilon n_0^{1/3}\right) \leq 2\varepsilon\sigma^2 + \mathbb{P}\left(N_n \notin [n_1, n_2]\right),$$

and by Goncharov's theorem this last term goes to zero as $n \to \infty$. Since ε can be made arbitrarily small, this completes the proof. □

For some additional details, we can check Diaconis and Tung's "Poisson approximation for large permutation groups." The fact we used in the proof above is the following:

Theorem 48

Suppose X_i have mean zero and finite variance. Then

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| > \lambda\right) \leq \frac{N\sigma^2}{\lambda^2}.$$

So the point is that if we've chosen our random sum in appropriate small-enough intervals, then we can get good enough bounds. And this argument is typical of what's called Anscombe's theorem, which says that $\frac{\sum_{i=1}^N X_i}{\sqrt{N}}$ converges to a normal random variable under appropriate conditions:

Theorem 49 (Anscombe)
Let $\{X_i\}_{i=1}^{\infty}$ be iid with mean zero and positive finite variance σ^2. Set $S_n = \sum_{i=1}^{n} X_i$. Now suppose that $\tau(n)$ are positive integer-valued random variables, and (there are other abstract or weaker conditions we can use instead) suppose that $\frac{\tau(n)}{n} \to \theta$ in probability for some finite θ. Then the randomly stopped sums $\frac{S_{\tau(n)}}{\sqrt{\tau(n)}}$ converge to $N(0, \sigma^2)$ as $n \to \infty$.

The point is that $\tau(n)$ can even be very dependent on the Xs (such as the number of times it's equal to zero) as long as they have the correct growth rate, and for a reference and many applications to topics like sequential analysis we can see Allan Gut's "Anscombe's theorem 60 years later."

Remark 50. *For some future research directions, the theorem in Theorem 47 takes k fixed, and it may be possible to also get a similar result for k growing with n. And similarly, it should be possible to take other groups besides $H = S_n$ as long as they are appropriately growing.*

Example 51
Professor Diaconis did a card trick in class which is basically about the hyperoctahedral group. The idea is to start off with $2n$ cards, arranged in "stay stack" (meaning that they start off as $1, 2, \cdots, n, n, \cdots, 1$, or any other arrangement where the top and bottom card pair up, the next top and next bottom card pair up, and so on, so that we get central symmetry).

It's then a math fact that we can ask to "shuffle the deck" in a variety of ways (which Professor Diaconis do) which all preserve the stay stack property. For example, if we deal the cards into any divisor (meaning we put them into $k|2n$ piles sequentially) and then pick them up from left to right or right to left, we still have central symmetry; similarly if we reverse shuffle (put cards alternating up and down in packets of size $j|n$, and then put all the up cards above all the down cards) or perfect shuffle, the property is still preserved.

From there, the next stage of the card trick is to do a **Monge shuffle**, which starts with the top card, puts the second card on top, the third card on the bottom, the fourth card on top, and so on. Instead of getting a pattern like 123456654321, this will yield a pattern like 123456123456 instead. Many shuffles preserve this pattern – in particular arbitrary random cuts. And so we can recover a matching pair of the cards at any point by taking the top card off, and (this last part is why Professor Diaconis used 12 cards in the trick) with 11 cards remaining, doing the "down-and-under" shuffle means the middle card comes out last.

Thus, it may be interesting to ask if we can find more shuffles that also preserve stay-stack (that is, what permutations preserve this property); the ones that Professor Diaconis did in class are the most natural to actually perform, but there are other ones. And note that the centrally symmetric permutations in S_{2n} form the group B_n that we described previously in this curse – we can swap a pair of matching cards or permute the pairs – but what's also interesting is that our Monge shuffle gives a correspondence of B_n with another copy of B_n. It turns out there's a third copy of B_n that's also involved; the standard copy of B_n in S_{2n} is the set of permutations where adjacent pairs add to $2n+1$ (so we can switch adjacent pairs or permute their order). And we can get between all of these with various operations (and all the copies of B_n are conjugate in S_{2n}); for example, the **milk shuffle** takes stay-stack to the standard arrangement.

Fact 52

If we have $2n$ cards and deal them into k piles (even if k does not divide $2n$), it turns out there is always a way of picking them up that preserves stay stack. For more along these lines, we can see the paper "The Magic of Charles Sanders Peirce" by Diaconis and Graham.

8 April 16, 2025

We've been explaining how to use generating functions to do probability, and last time we did this to get information about the cycle counts (showing that they converge to a normal distribution). We'll do a different kind of example today:

Example 53 (Cycles for wreath products)

Let $\Gamma \subseteq S_k$ and $H \subset S_n$ be subgroups, and let $G = \Gamma^n \rtimes H \subseteq S_{kn}$. Then an element of G has cycle type $\prod i^{a_i(\sigma)}$, and here's the question we want to ask: for $s \in G$ chosen uniformly, what is the joint distribution of $\{a_i\}_{i=1}^{kn}$?

We already know that if Γ is the identity and H is S_n, we have $G = S_n$ and we've shown that asymptotically the cycle counts $\{a_i\}_{i=1}^{n}$ converge (as a vector) to the distribution

$$\{A_i\}_{i=1}^{\infty}, \quad A_i \sim \text{Pois}\left(\frac{1}{i}\right) \text{ independent.}$$

Many properties can be read off of this fact (such as the length of the largest cycle), and we can understand the answer in general. (For more details, we may again see Diaconis and Tung's "Poisson approximation for large permutation groups.") First, we'll do some special cases for illustration:

Example 54

Let $\Gamma = S_3$ and $H = S_n$, so that $G = S_3^n \rtimes S_n$ is a subgroup of S_{3n}. If $s \in G_n$ is chosen uniformly and $\{a_i(s)\}_{i=1}^{3n}$ is the joint distribution of the number of i-cycles, then we have convergence of the vector $\{a_i(s)\}_{i=1}^{3n}$ to $\{A_i\}_{i=1}^{3n}$ as $n \to \infty$, where

$$A_i = \begin{cases} 3W_i + Z_i & i \equiv 1 \bmod 6, \\ 3W_i + Z_i + Z_{i/2} & i \equiv 2 \bmod 6, \\ 3W_i + Z_i + Y_i & i \equiv 3 \bmod 6, \\ 3W_i + Z_i + Z_{i/2} & i \equiv 4 \bmod 6, \\ 3W_i + Z_i & i \equiv 5 \bmod 6, \\ 3W_i + Z_i + Z_{i/2} + Y_i & i \equiv 0 \bmod 6, \end{cases}$$

where $\{W_i, Z_i, Y_i\}_{i=1}^{\infty}$ are all independent and the distributions of W_i, Z_i, Y_i are Poissons with parameters $\frac{1}{6i}, \frac{1}{2i}, \frac{1}{i}$, respectively.

In particular, note that the A_is aren't independent, because both of

$$A_1 = 3W_1 + Z_1, \quad A_2 = 3W_2 + Z_2 + Z_1$$

have dependence through Z_1. But many of the cycle counts are independent (for example, $A_1, A_3, A_5, A_7, \cdots$ have no

common terms, and so do $A_j, A_{j+1}, \cdots, A_{2j-1}$ for any fixed j), and we can still figure out questions like the distribution of cycle statistics.

Example 55

Let $\Gamma = C_k$ and $H = S_n$ (so in particular if $k = 2$ this yields the hyperoctahedral group). Then if s is chosen uniformly from $G_n = C_k^n \rtimes S_n$, then the cycle counts $\{a_i(s)\}$ converge to $\{A_i\}$, where

$$A_i = \sum_{\ell | (i,k)} \frac{k}{\ell} Y_{i,\ell}, \quad Y_{i\ell} \sim \text{Pois}\left(\frac{\ell \phi(\ell)}{ki}\right),$$

where (i,k) denotes the gcd of i,k and $\phi(\ell)$ is the Euler totient function.

In particular, this time all of the A_is are independent. Notice that in all cases we've described here, the A_is are **compound Poisson** – let's explain what that means.

Definition 56

Let $\{X_i\}_{i=1}^{\infty}$ be iid integer-valued random variables with $\mathbb{P}(X_i = j) = \theta_j$ for nonnegative constants $\theta_1, \theta_2 \cdots$ summing to 1. Let N be Poisson with parameter λ. Then a **compound Poisson** with parameters (λ, θ_j) is a variable of the form $W = \sum_{i=1}^{N} X_i$.

In particular, these Ws are infinitely divisible for any parameters (so for any k, we can find k iid random variables such that adding them together is equal in distribution to W). To explain what that means, the central limit theorem says that we often get a bell-shaped curve when adding together iid copies of a random variable, and we might ask for all the limit laws (that is, random variables where adding n independent copies, subtracting off a_n, and dividing by b_n yields some limiting distribution). For integer-valued random variables, this turns out to be exactly the infinitely divisible distributions, so it's indeed a natural question to think about.

The sum of Poissons of parameter λ_1, λ_2 is again Poisson with parameter $\lambda_1 + \lambda_2$, so we get infinite divisibility by just replacing λ with $\frac{\lambda}{k}$. And it turns out that for **any** infinitely divisible law, we can write it in this way for some choice of (λ, θ_j) – we can see Feller volume 1 for this.

Proposition 57

The compound Poisson can also equivalently be described as follows: let Y_j (for $j \geq 1$) be independent Poissons of parameter $\lambda\theta_j$, and define $W' = \sum_{j=1}^{\infty} jY_j$. Then W' has the same law as W.

Proof. Consider the generating function

$$\mathbb{E}[z^W] = \mathbb{E}\left[\mathbb{E}[z^W | N = j]\right] = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \phi(z)^j,$$

where $\phi(z) = \mathbb{E}[z^{X_1}] = \sum \theta_J z^j$ is the generating function for an individual part. We can further simplify this by power series manipulation to

$$e^{-\lambda} e^{\lambda \phi(z)} = e^{-\lambda} e^{\lambda \sum_{j=1}^{\infty} \theta_j z^j} = e^{\lambda \sum_{j=1}^{\infty} \theta_j (z^j - 1)} = \prod_{j=1}^{\infty} e^{\lambda \theta_j (z^j - 1)},$$

and this is exactly the generating function of Y' we have described, since $\mathbb{E}[z^{jY_j}] = e^{\lambda(z^j - 1)}$ for a Poisson of parameter λ. □

This problem was on the qualifying exam for statistics graduate students last year, but it's important for us too! Compound Poissons occur in various places, and one good reference (which uses Stein's method) is Barbour and Chryssaphinou's "Compound Poisson approximation: a user's guide" (which shows various places where limit theorems come up).

Theorem 58

Fix k, let $\Gamma \subseteq S_k$ and $G_n = \Gamma^n \rtimes S_n$, and let G_∞ be the union of the G_ns. For every $0 < t < 1$, define the probability measure U_t on G_∞ by first picking N so that $\mathbb{P}(N = n) = (1-t)t^n$ (that is, sample from a geometric distribution) and then sample σ uniformly from G_n. Then under U_t, the vector of cycle counts $\{a_i(\sigma)\}_{i=1}^\infty$ (this is always eventually zero) is equal in distribution to $\{A_i\}_{i=1}^\infty$, where

$$A_i = \sum_{\substack{\lambda \vdash k \\ j \cdot \ell = i}} a_j(\lambda) Z_{\ell,\lambda}.$$

Here we're doing a sum over partitions λ and integers j, ℓ (but $\ell = \frac{i}{j}$), where $Z_{\ell,\lambda}$ is Poisson with parameter $\frac{t^\ell}{\ell} p_\Gamma(\lambda)$ (and $p_\Gamma(\lambda)$ is the probability that a uniform random permutation of $\Gamma \subseteq S_k$ has cycle type λ), $a_j(\lambda)$ is the number of parts of size j in λ, and all $Z_{\ell,\lambda}$s are independent.

If we take $t \to 1$, then we get a limit theorem which tells us the distribution that G_n converges to for a fixed n. But the theorem here is exact, and it shows us indeed that the A_is are generally not independent.

Proof sketch. Recall that if X, Y, Z are independent Poisson random variables with parameter λ, μ, ν, we have

$$\mathbb{E}[z^{jX}] = e^{\lambda(z^j - 1)},$$

and if $A = jX + \ell Z$ and $B = kY + \ell' Z$ we have the joint generating function of these coupled linear combinations

$$\mathbb{E}[x^A y^B] = e^{\lambda(x^j - 1) + \mu(y^k - 1) + \nu(x^\ell y^{\ell'} - 1)}.$$

The point is that we can build fancy polynomials on the right-hand side by appropriately coupling. In general, **if** for every subset $S \subseteq [n]$ we have independent Poisson random variables X_S of parameter λ_S, and we have constants $C_S^i \in \mathbb{N}$ and define the general integer linear combinations of Poissons

$$W_i = \sum_{S \subseteq [n]} C_S^i X_S,$$

then we get the joint generating function

$$\mathbb{E}\left[\prod_{i=1}^N x_i^{W_i}\right] = \exp\left(\sum_{S \subseteq [n]} \lambda_S \left(\prod_{i=1}^N x_i^{C_S^i} - 1\right)\right).$$

What we can thus do is return to our cycle index story and consider the joint generating function of our random variables

$$\sum_{n=0}^\infty Z_{G_n}(x_1, \cdots, x_{kn})(1-t)t^n = \exp\left(\sum_{a=1}^\infty \frac{t^a}{a}\left(Z_\Gamma(x_a, x_{2a}, \cdots, x_{ka}) - 1\right)\right).$$

Now if we expand out the exponent on the right-hand side, we get

$$\exp\left(\sum_{a=1}^\infty \frac{t^a}{a} \sum_{\lambda \vdash k} P_\Gamma(\lambda)\left(\prod_{b=1}^k x_{ab}^{a_b(\lambda)} - 1\right)\right),$$

so if we switch the order of summation we get

$$\exp\left(\sum_{\lambda\vdash k}\sum_{a=1}^{\infty}\frac{t^a}{a}P_T(\lambda)\left(\prod_{b=1}^{k}x_{ab}^{a_b(\lambda)}-1\right)\right)$$

and this exponent is the log generating function of compound Poissons – picking out the terms where $x_{ab}=x_i$ to get the distribution of a particular cycle count yields the result. □

So this uses the plethysm theorem, randomization, and recognizing the generating function of compound Poissons. That gives us results for randomized n, but now we might be curious about finite n as well:

Theorem 59

For $\sigma\in G_n$, $\{a_i(\sigma)\}_{i=1}^{kn}$ converges in distribution to $\{\tilde{A}_i\}_{i=1}^{\infty}$, where $\tilde{A}_i$ is the same as A_i but with $t=1$.

This can be done analytically using the generating functions, or it can be done with coupling – we can get an error term, showing that if $b<n$ is fixed, then

$$\left|\left|\mathcal{L}(a_i:1\le i\le b)-\mathcal{L}(\tilde{A}_i:1\le i\le b)\right|\right|_{\text{TV}}\le\frac{2b}{n}.$$

So these generating functions do have probabilistic content, and there's much more we can do with them as well. We'll do one more example next time and then move to more "classical" Polya theory from there!

9 April 18, 2025

Today's topic is **product actions** – we'll consider $S_k\times S_n$ acting on $[k]\times[n]$ via

$$(i,j)^{(\sigma,\tau)}=(\sigma(i),\tau(j)).$$

For example if $k=4,n=13$, we just picture a normal deck of cards with the usual suits and values, laid out in a 4×13 grid. We're then allowed to permute the rows and columns in any way we'd like – rows will always be constant on suit and columns will always be constant on value, but those are the only constraints. We might then ask about the cycle counts of the resulting permutation on kn elements.

There are two reasons we might study this (beyond the fact that Professor Diaconis was told when working on Burnside processes that he probably couldn't do it):

- There's a theory of group actions which is less obvious than we might think, and understanding product actions here is worthwhile.
- We get surprising limit theorems of a kind that's different from before.

The idea is that if (i,j) is fixed by (σ,τ), then we must have $\sigma(i)=i$ and $\tau(j)=j$. therefore $a_1(\sigma,\tau)=a_1(\sigma)a_1(\tau)$ (we fix some number of rows and columns, and the cells in those rows and columns are exactly the fixed points). We know that for k,n large each of $a_1(\sigma)$ and $a_1(\tau)$ converge to Poisson(1), and these are independent, so

$$a_1(\sigma,\tau)\to XY,\quad X,Y\text{ iid Poisson}(1).$$

Observe that the product is not infinitely divisible and hence is not compound Poisson, but it's still nice. We'll write

down the general formula, but to help illustrate it we'll also note that

$$a_2(\sigma,\tau) \to Y_2X_1 + (Y_1+2Y_2)X_2, \quad X_i, Y_i \text{ Poisson}\left(\frac{1}{i}\right) \text{ all independent.}$$

This is a special case of groups acting on product sets:

Proposition 60 (Polya)
The cycle index polynomial can be written as

$$Z_{S_k\times S_n}(x_1,\cdots,x_{kn}) = \sum_{\substack{\lambda\vdash k\\ \mu\vdash n}} \frac{1}{Z_\lambda Z_\mu}\prod_{i,j} x_{[i,j]}^{(i,j)a_i(\lambda)a_j(\mu)},$$

where $(i,j)=\gcd(i,j)$ and $[i,j]=\operatorname{lcm}(i,j)$.

For example, we can check that

$$Z_{S_2\times S_3}(x_1,x_2,x_3,x_4,x_5,x_6) = \frac{1}{12}\left(x_1^6+3x_1^2x_2^2+2x_3^2+4x_2^3+2x_6\right).$$

This implies the following:

Theorem 61
Pick (σ,τ) uniformly in $S_k\times S_n$. Then for k,n large, the vector of ℓ-cycles $\{a_\ell(\sigma,\tau)\}_{i=1}^{kn}$ converges to a limiting vector $\{A_i\}_{i=1}^{\infty}$, where

$$A_i = \sum_{a|i} X_a \sum_{j:[j,a]=i}(j,a)Y_j,$$

where $X_i, Y_i \sim \text{Poisson}\left(\frac{1}{i}\right)$ are all independent.

So the worst we do is multiply together terms with two Poissons. The proof just involves randomizing the parameter as we've already done, but let's see it worked out:

Proof. We'll use Proposition 60, as well as Polya's cycle index theorem. First fix k and randomize n; we have

$$\sum_{n=0}^{\infty} Z_{S_k\times S_n}(x_1,\cdots,x_{kn})(1-t)t^n,$$

and now when we do the double sum of Proposition 60 over λ,μ, we can rewrite this sum using Polya's theorem one λ at a time as

$$\sum_{\lambda\vdash k}\frac{1}{Z_\lambda}\exp\left(\sum_{a=1}^{\infty}\frac{t^a}{a}s_{\lambda,a}-1\right), \quad \text{where } s_{\lambda,a}=\prod_{i=1}^{k}x_{[a,i]}^{(a,i)a_i(\lambda)}.$$

This gives us the joint distribution of everything, but we can consider the marginal distribution of just a_ℓ. For this, we set $x_\ell = x$ and all other x_is to 1, so for fixed λ, a the expression for $s_{\lambda,a}$ reduces to

$$s_{\lambda,a} = x^{n_\ell(a,\lambda)}, \quad n_\ell(a,\lambda) = \sum_{i:[a,i]=\ell}(a,i)a_i(\lambda),$$

as long as $a|\ell$ (since otherwise there won't be any terms at all). Therefore

$$\sum \mathbb{E}_{S_k\times S_n}\left[x^{a_\ell(\sigma,\tau)}\right](1-t)t^n = \sum_{\lambda\vdash k}\frac{1}{Z_\lambda}\exp\left(\sum_{a|\ell}\frac{t^a}{a}x^{n_\ell(a,\lambda)}\right)$$

is the generating function of a compound Poisson $\sum_{a|\ell} X_a n_\ell(d, \lambda)$ with $X_a \sim \text{Pois}\left(\frac{t^a}{a}\right)$; letting $t \to 1$ gives us part of the marginal distribution we claimed. Then sending $k \to \infty$, we see that $a_i(\sigma) \to \text{Pois}\left(\frac{1}{i}\right)$. □

Fact 62

We can also do a finite version of this via a coupling proof: if we let $f(n)$ be any function growing to infinity, μ is the distribution of the first b cycles under $S_{f(n)} \times s_n$, and ν is the limit walk, then $||\mu - \nu|| \leq \frac{2b}{n} + \frac{2b}{f(n)}$.

The point is to see that knowing the cycle indices and Polya's theorem can lead to funny limit theorems in various cases.

Remark 63. *This proposition also generalizes by saying that G_1 acts on $[k]$ and G_2 acts on $[n]$ (rather than just using S_k and S_n); we get a similar formula for the cycle index $Z_{G_1 \times G_2}(x_1, \cdots, x_{kn})$ which involves terms of the same form with lcms and gcds. And this works for products with more than two terms as well, resulting in some funny algebra. For references on this, we can see Wei and Xu's "Cycle index of direct product of permutation groups and number of equivalence classes of subsets of Z_v;" this was a paper studying* **difference sets**, *which are used for constructing arrays of combinatorial designs (which are collections of subsets with equal incidences of pairs, for example). Specifically, this paper studies $x \mapsto ax + b \mod v$ (for a, b relatively prime) acting on k-sets of $\{0, 1, \cdots, v-1\}$, but it does so using Polya theory.*

The point is that if we have a cycle index, we can maybe do something with it, and the question then is "what cycle indices do we know?". We started with Z_{C_n} and Z_{S_n}; we can also do the dihedral group Z_{D_n} (useful for chemistry, since we might want to classify something like the benzene molecule under the usual symmetries) and the alternating group Z_{A_n} with some more work. We then also did wreath products and product groups in these last few lectures. So an interesting research problem would be to do semidirect products (where we know the cycle index of the quotient and normal subgroup) – it's probably doable and hasn't really been done yet. And to end on a positive note, there's an interesting paper "Cycle indices of linear, affine, and projective groups" by Fripertinger which does cases like $GL_n(\mathbb{F}_q)$ or the corresponding affine group. (This is another world where Polya theory is applied, and this group is studying linear codes and might want to do enumeration.)

We'll move topics now – next we'll talk about the Burnside process, understanding how to understand orbits via simulation.

10 April 21, 2025

We've been doing something in the flavor of generating functions and asymptotics, and so today we'll move to something different.

Example 64

Let $\mathfrak{X}$ be a finite set and G a finite group acting on $\mathfrak{X}$. Under this action, $\mathfrak{X}$ is split into orbits, and as usual we are curious about enumeration, sizes, typical behavior, and other properties of those orbits.

In particular, we might be curious how we can pick an orbit uniformly at random; an answer to this last question often gets us some kind of answer to the others via sampling.

Definition 65

The **Burnside process** is a Markov chain on $\mathfrak{X}$ defined as follows:

- From $x \in \mathfrak{X}$, pick $s \in G$ uniformly from the set $G_x = \{s : x^s = x\}$. (This set always contains the identity element, so it is nonempty.)
- From $s \in G$, pick $y \in \mathfrak{X}$ uniformly from the set $\mathfrak{X}_s = \{y : y^s = y\}$. (This set always contains x, so it is nonempty.)

One step of the Markov chain then takes us from x to y.

We can write down explicitly the transition matrix of this chain

$$K(x,y) = \frac{1}{|G_x|} \sum_{s \in G_x \cap G_y} \frac{1}{|\mathfrak{X}_s|},$$

since we have to pick s fixing both x and y. We have $K(x,y) \geq 0$ for all x, y and algebraically can check that $\sum_y K(x,y) = 1$, so this is indeed a Markov kernel.

Theorem 66

The Burnside process described above is ergodic (that is, connected and aperiodic), and its unique stationary distribution is

$$\pi(x) = \frac{1}{z|\mathcal{O}_x|},$$

where z is the number of orbits under the group action.

Here, the stationary distribution is the "long-term equilibrium distribution," and algebraically it's described by saying that $\sum_x \pi(x)K(x,y) = \pi(y)$. Furthermore, it turns out that (π, K) is **reversible**, meaning that $\pi(x)K(x,y) = \pi(y)K(y,x)$ for all $x, y \in \mathfrak{X}$ – this condition is often called **detailed balance**, and that is useful for various things.

Proof. The fact that $K(x,y) > 0$ for all x, y (since we can always go from x to the identity to y) shows ergodicity (no parity problems because we have a positive probability of staying at x). By the orbit-stabilizer theorem, $|\mathcal{O}_X| = \frac{|G|}{|G_x|}$, so we can rewrite $\pi(x) = \frac{1}{z}\frac{1}{|\mathcal{O}_x|} = \frac{|G_x|}{z|G|}$ and then verify the condition for reversibility:

$$\pi(x)K(x,y) = \frac{1}{z}\frac{|G_x|}{|G|} \cdot \frac{1}{|G_x|} \sum_{s \in G_x \cap G_y} \frac{1}{|\mathfrak{X}_s|} = \frac{1}{z|G|} \sum_{s \in G_x \cap G_y} \frac{1}{|\mathfrak{X}_s|}$$

and this last expression is symmetric in x, y and thus is also equal to $\pi(y)K(y,x)$. Finally, reversibility implies stationarity because

$$\sum_x \pi(x)K(x,y) = \sum_x \pi(y)K(y,x) = \pi(y)$$

with the last step coming from $\sum K(y,x) = 1$ for a stochastic matrix. And $\frac{1}{z}$ provides the correct normalizing constant, because summing $\frac{1}{|\mathcal{O}_x|}$ over any orbit yields 1 and we want the total sum over all orbits to be 1. □

Thus if we can run the Burnside process and keep track of the orbit we're in at each step, this yields another process. But the probability of winding up at any point in the orbit is the same, so we get a random process which ends up uniformly distributed over all possibilities.

Example 67

Let $\mathfrak{X} = C_2^n$ and $G = S_n$, where G acts on $\mathfrak{X}$ by permuting coordinates. The orbits of this action are the level sets

$$\mathcal{O}_i = \{x : |x| = i\},$$

where $|x|$ is the number of ones in the binary n-tuple x.

Thus the all-zeros vector is its own orbit, but in general the size of the orbit $\mathcal{O}_i$ is $\binom{n}{i}$ (which widely varies). We'll think through how we would run the Burnside process in this case:

- We start with a binary n-tuple x and want to pick a permutation σ that fixes it. In order to have $x^\sigma = x$, we must choose from the isotropy subgroup
$$G_x = S_i \times S_{n-i},$$
where S_i permutes the locations of the ones and S_{n-i} permutes the locations of the zeros. (This is easy to actually do, so sampling here is easy.)
- Now from a permutation $\sigma \in S_n$, we want to pick a binary n-tuple y such that y is fixed by σ. We do this by writing σ in cycle notation, and we must pick y to be constant on each cycle. Thus we independently label each cycle to be 1 or 0 with equal probability and label all of those positions with the corresponding label.

It's not always so easy to do these two steps, but it is often doable. If we similarly let $\mathfrak{X} = C_k^n$ and $G = S_n$ act by permuting coordinates again, then the orbits are now of the form

$$\{x : x \text{ has } n_1 \text{ ones}, n_2 \text{ twos}, \cdots, n_k \ k\text{s}\}$$

and are thus indexed by tuples $(n_1, \cdots, n_k)$ with $0 \leq n_i \leq n$ for all i and $\sum_{i=1}^k n_i = n$. The uniform distribution on all such tuples is the **Bose-Einstein distribution**, and we have (by stars and bars)

$$\mathbb{P}_{\text{BE}}(n_1, \cdots, n_k) = \frac{1}{\binom{n+k-1}{k-1}}.$$

There are lots of stories we can tell about this (and we'll do so in a few lectures), but we can confirm that when $k = 2$ this reduces to the uniform distribution of assigning $(i, n-i)$ probability $\frac{1}{n+1}$ for all $0 \leq i \leq n$. And the point is that this process gives us dynamics which have Bose-Einstein as a stationary distribution, but this is rather different than something like the Metropolis algorithm which makes much more local moves.

Remark 68. *The Burnside process is then carried out in exactly the same way as for (C_2^n, S_n): we permute among all coordinates with the same value to get σ, and then we write it in cycle notation and independently and uniformly pick a label on each cycle to get y.*

We'll see many new examples as we proceed, but what we really want to do is sample directly from the orbit process. In general if we group our states into lumps and only report what lump we're in at each stage, that isn't going to be a Markov chain. (For example, consider simple random walk on the n-point circle where we lump the left and right half. Then it's not true that "the future depends on the past through the present:" seeing left a bunch of times in a row means we're more likely to see another left, but seeing alternating lefts and rights does not.)

However, in this case we have some additional symmetry which makes things nicer:

Lemma 69

The Burnside process lumped to orbits is again a Markov chain with uniform stationary distribution.

Proof. We use Dynkin's criterion, which says the following:

Theorem 70 (Dynkin's criterion)
Let $K(x,y)$ be any Markov chain on $\mathfrak{X}$, and let $\mathcal{O}_1 \cup \cdots \cup \mathcal{O}_\ell$ be any partition of the space. Let $f : \mathfrak{X} \to [\ell]$ be the function where $f(x) = i$ if $x \in \mathcal{O}_i$. We have a Markov chain $X_1, X_2, \cdots$ on $\mathfrak{X}$, and we have a corresponding process $Y_i = f(X_i)$ on $\{1, \cdots, \ell\}$. Then Y_i is a Markov chain for **every** starting distribution of X_1 if and only if for every $\mathcal{O}, \mathcal{O}'$ and every $x, y \in \mathcal{O}$, we have $K(x, \mathcal{O}') = K(y, \mathcal{O}')$.

We can see a proof of this for instance in Kemeny and Snell's "Finite Markov Chains," or Pang's survey paper "Lumpings of Algebraic Markov Chains arise from Subquotients." So we can verify this condition for the Burnside process: for any $\mathcal{O}\mathcal{O}'$ and $x, y \in \mathcal{O}$, we know that $y = x^t$ for some $t \in G$ (because they are in the same orbit of the group action). We can then check that

$$G_x^t = t^{-1} G_x t = G_{x^t},$$

so in particular $|G_x| = |G_{x^s}| = |G_y|$. Furthermore, we have for any x, z that

$$s \in G_x \cap G_z \iff s^t \in G_{x^t} \cap G_{z^t}$$

and also that $|\mathfrak{X}_s| = |\mathfrak{X}_{s^t}|$ (by just writing out definitions), and so

$$\begin{aligned} K(x,z) &= \frac{1}{|G_x|} \sum_{s \in G_x \cap G_z} \frac{1}{|\mathfrak{X}_s|} \\ &= \frac{1}{|G_{x^t}|} \sum_{s^t \in G_{x^t} \cap G_{z^t}} \frac{1}{|\mathfrak{X}_{s^t}|}, \end{aligned}$$

and now writing $u = s^t$ this simplifies to

$$\frac{1}{|G_y|} \sum_{u \in G_y \cap G_{z^t}} \frac{1}{|\mathfrak{X}_u|} = K(y, z^t).$$

But if $K(x,z) = K(y,z^t)$, then summing over the whole orbit $z \in \mathcal{O}'$ yields $K(x,\mathcal{O}') = K(y,\mathcal{O}')$, as desired.

So the orbit chain is a Markov chain, and the transition probabilities $K(\mathcal{O},\mathcal{O}')$ are given by $K(x,\mathcal{O}')$ for **any** choice of $x \in \mathcal{O}$. And for any lumped chain satisfying Dynkin's criterion we always have

$$\pi(\mathcal{O}) = \sum_{x \in \mathcal{O}} \pi(x),$$

and since this sum is $\frac{1}{z|\mathcal{O}_x|}$ over all $x \in \mathcal{O}$ it exactly evaluates to $\frac{1}{z}$, which is indeed uniform on orbits as desired. □

This lumping business is important in general, and we'll do one example where we can derive the lumped chain:

Example 71
Returning to the (C_2^n, S_n) example, we can understand the lumped chain for the Burnside process and write down the resulting Markov chain on $\{0, 1, \cdots, n\}$.

To describe it, we need to know about the **discrete arcsine law** – this is a probability distribution defined by

$$\alpha_k^n = \frac{\binom{2k}{k}\binom{2n-2k}{n-k}}{2^{2n}}, \quad 0 \le k \le n.$$

Graphed as a function of k, this is largest near 0 and n and smallest at $\frac{n}{2}$, and if we rescale time and space this converges to the smooth curve $\frac{1}{\pi\sqrt{x(1-x)}}$, which is the integral of arcsine. For more information on this see Feller volume 1, chapter 3, but if we do coin flipping $2n$ times and let K be the last time that the number of heads is equal to the number of tails (which is always even) then $K = 2k$ with k following the discrete arcsine distribution. (And as a corollary, 20 percent of the time, either heads or tails stays ahead for 98 percent of the time when do a series of fair coin flips.)

11 April 23, 2025

Today's discussion will be on the **lumped Burnside process**. Specifically, we were studying (C_2^n, S_n) with S_n permuting the coordinates of the binary n-tuples, and we showed that the Burnside process lumps to a Markov chain on the orbits $\{\mathcal{O}_0, \cdots, \mathcal{O}_n\}$ (where $\mathcal{O}_i$ is the set of n-tuples with exactly i ones); in particular this yields a chain with uniform stationary distribution on the $\mathcal{O}_i$s. Thus, we might want to write down an expression for $K(i,j)$ (in particular so that we can run the lumped process). We can notice some obvious symmetries:

Theorem 72

We have $K(i,j) = K(j,i) = K(i,n-j) = K(n-i,j)$ for all $0 \le i,j \le n$. Also, the transition probabilities

$$K(0,k) = \alpha_k^n = \frac{\binom{2k}{k}\binom{2n-2k}{n-k}}{2^{2n}}.$$

are given by the discrete arcsine distribution, and in general

$$K(j,k) = \sum_{\ell} \alpha_\ell^j \alpha_{k-\ell}^{n-j}$$

where the sum is over all indices $(j+k-n)_+ \le \ell \le j \wedge k$ which make the terms nonnegative.

For example when $n = 2$, we have

$$K(0,0) = \frac{6}{16}, \quad K(0,1) = \frac{4}{16}, \quad K(0,2) = \frac{6}{16}.$$

Proof. The identities in the first sentence are clear because in the description of the Burnside process we can "flip the roles of 0s and 1s" and the chain is reversible. Furthermore, the general identity $K(j,k)$ follows from the case $K(0,k)$, because in order to end up with k ones we must get ℓ ones from the j coordinates which started as ones, and we must get the remaining $k-\ell$ ones from the $n-j$ coordinates which started as zeros, summing over all possible ℓ. Thus we just need to prove the formula for $K(0,k)$.

For this, let's first do the case $K(0,0)$. We first pick a random permutation $\sigma \in S_n$ and write it as a product of cycles; in order to end up with no ones, each cycle must flip an independent coin and not flip heads (aka must be labeled with 1). Thus

$$K(0,0) = \frac{1}{n!}\sum_{\sigma \in S_n} \left(\frac{1}{2}\right)^{a_1+\cdots+a_n} = Z_{S_n}\left(\frac{1}{2}, \cdots, \frac{1}{2}\right).$$

We know by Polya's theorem that

$$\boxed{\sum_{n=0}^{\infty} Z_{S_n}(x_1, \cdots, x_n)t^n = \exp\left(\sum \frac{t^i}{i} x_i\right)} \implies \sum_{n=0}^{\infty} K_n(0,0)t^n = \exp\left(\sum \frac{t^i}{i}\right)^{1/2} = \frac{1}{\sqrt{1-t}},$$

and now by the binomial theorem we can expand out the right-hand side as $\sum_{n=0}^{\infty} \binom{-1/2}{n}(-t)^n$, so that the coefficient of t^n is

$$K_n(0,0) = \frac{\frac{1}{2} \cdot \frac{3}{2} \cdot \cdots \cdot \frac{2n-1}{2}}{n!} = \frac{\binom{2n}{n}}{2^{2n}},$$

which agrees with what we expect. Similarly, we can only end up with one 1 if exactly one fixed point ends up heads:

$$K_n(0,1) = \sum_{\sigma \in S_n} a_1 \left(\frac{1}{2}\right)^{a_1 + \cdots + a_n}.$$

We get this by differentiating the boxed expression once in x_1 (yielding an a_1 factor) and then setting all x_i to $\frac{1}{2}$; this yields a right-hand side of $\frac{t}{\sqrt{1-t}}$, and we can compare coefficients again and it will work out to α_1^n. To get $K_n(0,k)$ in general, we can write it out in terms of cycles (the sum of all the possible ways to get cycle lengths adding to k) but we can always evaluate that with an appropriate derivative of the cycle index polynomial and the same combinatorics we've been doing. □

The details can be found in Professor Diaconis' paper "Analysis of a Bose-Einstein Markov chain." So we have our discrete arcsine distribution, and we may ask how we can sample k from α_k^n in an efficient way.

Proposition 73

Suppose we pick $\theta \in (0,1)$ from the continuous beta distribution $B(\frac{1}{2}, \frac{1}{2}, x) = \frac{1}{\pi\sqrt{x(1-x)}}$. Then sampling $k \sim \text{Bin}(n, \theta)$ will yield the discrete arcsine distribution. In other words,

$$\int_0^1 \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\pi\sqrt{\theta(1-\theta)}}\theta = \alpha_k^n.$$

Since it's easy to generate α_k^n, it's therefore easy to run $K(i,j)$, which is a convolution of two arcsine laws (coming from the coordinates on 0 and the coordinates on 1).

Remark 74. *By checking binomial coefficients, we know that $K(0,\ell)$ is smallest when $\ell = \frac{n}{2}$, and we can check using bounds on central binomial coefficients that $K(0, \lfloor \frac{n}{2} \rfloor) \geq \frac{1}{\pi n}$. This will be useful in a second.*

Remark 75. *The arcsine distribution $B(\frac{1}{2}, \frac{1}{2}, x)$ comes up in a variety of problems (for example related to Brownian motion). One that we'll state here goes as follows: make a process on $(0,1)$ by starting u_1 uniform on the interval, picking either "left" or "right" at random with probability $\frac{1}{2}$, and picking u_2 uniformly on that interval. Then we pick u_3 either left or right of u_2 at random and so on, but never go past a previous point so we get a nested sequence of intervals. It turns out that the limit is here is $B(\frac{1}{2}, \frac{1}{2})$, and an interesting question is whether there is a finite version of this on $\{0, 1, \cdots, n\}$ corresponding to the discrete arcsine. For more, we can see Professor Diaconis' paper with Kemperman "Some New Results for Dirichlet Priors," which is motivated by the* **Markov moment problem**.

So we have the lumped Burnside process described, and we're now going to try to answer questions about rates of convergence. For $\mathfrak{X}$ a finite set and K an ergodic Markov kernel with stationary distribution $\pi(x)$, we may be curious how long it takes for the chain to be close to stationary. Mathematically, given $\varepsilon > 0$, we can ask how large ℓ needs to be for

$$||K_x^\ell - \pi||_{\text{TV}} < \varepsilon,$$

where the **total variation** $||\cdot||_{\text{TV}}$ is defined by

$$||K_x^\ell - \pi||_{\text{TV}} = \frac{1}{2}\sum_y |K^\ell(x,y) - \pi(y)| = \max_{A \subseteq \mathfrak{X}} |K^\ell(x,A) - \pi(A)|$$

(that is, each row of matrix powers of K tends to the vector π, and we want to look at the ℓ^1 norm of the difference). We can bound rates of convergence in many ways, but today we'll do the **Doeblin condition**:

Theorem 76

Suppose there exists a fixed integer a and some $0 < c < 1$ such that $K^a(x, y) \geq c\pi(y)$ for all x, y. Then

$$||K_x^\ell - \pi||_{\mathrm{TV}} \leq (1-c)^{\ell/a}.$$

This tells us that total variation decays exponentially, and this is useful for chains which "go a long way in one step" so that we can make a small. But it's not so useful for example for simple random walk on an n-point circle, which would need $a = n^2$ to get c not too small (say $\frac{1}{\sqrt{n}}$) – this is not the correct bound for rate of convergence in this chain.

Either way, for our lumped Burnside chain on (C_2^n, S_n), we have $K(0, j) \geq \frac{1}{\pi n} \approx \frac{1}{\pi}\pi(j)$, so we can prove by induction that even taking $a = 1$, $K(i, j) \geq \frac{1}{\pi}\pi(j) + O\left(\frac{1}{n^2}\right)$. Thus the Doeblin condition tells us that

$$||K_0^\ell - \pi||_{\mathrm{TV}} \lesssim \left(1 - \frac{1}{\pi}\right)^\ell,$$

which in particular tells us that a bounded number of steps is sufficient for convergence. Therefore, a bounded number of steps is sufficient for convergence even on the unlumped chain (C_2^n, S_n) if we start from the all-zeros state, or if we start uniformly on any orbit. However, there are starting points on the full (unlumped) $\mathfrak{X}$ chain where it does take longer, and we'll talk about that later.

This is the simplest example of a Burnside process, but there are close generalizations that are interesting:

Example 77 (Bose-Einstein statistics)

Consider the Burnside process on (C_k^n, S_n) (so that we have n-tuples taking one of k values instead of just binary ones). This is carried out in exactly the same way as for (C_2^n, S_n), except that we pick uniform permutations among each label and then label the resulting permutation uniformly with one of the k possibilities. We then label our orbits by tuples of nonnegative integers $(n_0, \cdots, n_{k-1})$ with $\sum n_i = n$, and the Burnside process is uniform on orbits. Therefore we get the Bose-Einstein distribution $\mathbb{P}_{\mathrm{BE}}(n_1, \cdots, n_k) = \frac{1}{\binom{n+k-1}{k-1}}$ on orbits.

This is a famous measure because of Bose-Einstein condensation, and what's important is that it's a very different measure than iid assigning a value to each tuple. (For example for $k = 2$, Bose-Einstein is uniform on orbits, but iid assignments is binomial and thus very sharply peaked near $\frac{n}{2}$.) So the question we care about is "what does a typical Bose-Einstein configuration look like?," and we'll show how to answer those next time.

12 April 25, 2025

We're trying to use as a unifying theme the question of understanding typical orbits of group actions, and today we'll do this in a specific case where the math is interesting (continuing the example of S_n acting on C_k^n). We know from what we were doing that orbits are indexed by the counts $(n_1, \cdots, n_k)$ of each element of the alphabet, and that the stationary distribution is uniform over all possible counts (the Bose-Einstein distribution).

Intuitively, we can think of this as saying that we drop n unlabeled balls into k boxes in a way where all configurations are equally likely; this is very different from dropping each ball in at random. This does come up in lots of places,

and we'll discuss some of them today – the question we're interested in is how to count the number of empty cells, maximum cell count, and so on.

- First of all, we saw this as the stationary distribution of a Burnside process that we care about.
- A second appearance comes from physics – if we have n sparse bosons (Higgs, photons, lead, certain isotopes of rubidium – this is just one class of particles that's different from fermions), and they need to go onto k shells, people initially thought that each particle is independent of the others and so they'll fall independently like Maxwell-Boltzmann
$$\mathbb{P}(n_1, \cdots, n_k) = \frac{1}{k^n} \binom{n}{n_1, \cdots, n_k}.$$
But Bose and Einstein were corresponding about some anomaly, and it turned out experimentally that indeed this is not the right distribution because bosons are indistinguishable. (We can check Professor Diaconis' paper with Chatterjee, "Fluctuations of the Bose-Einstein condensate," for more details.)
- This comes up in Bayesian statistics as well, and that's what we'll discuss in more detail below.

Example 78

In probability, we have the **birthday problem**, and in the classical case we assume everyone's birthday is iid uniformly distributed – this is saying that we drop n balls (kids) into k boxes (birthdays) and want the probability that two balls fall in the same box, which is 1 minus the probability that all balls fall in distinct boxes (which is $\left(1-\frac{1}{k}\right)\left(1-\frac{2}{k}\right)\cdots\left(1-\frac{n-1}{k}\right)$).

We can also write this as
$$\exp\left(\sum_{j=1}^{n-1} \log\left(1-\frac{j}{k}\right)\right) \sim \exp\left(-\sum_{j=1}^{n-1}\frac{j}{k}\right) = \exp\left(-\binom{n}{2}/k\right),$$
so we can choose n accordingly to get whatever probability we want – the answer turns out to be roughly $n = 1.2\sqrt{k}$, so we need about 23 people.

But it's probably not true that the distribution is uniform (for example roughly 20 percent fewer births happen on weekends than weekdays, seasonal effects, etc.). So it makes sense to instead use a multinomial distribution $\binom{n}{n_1,\cdots,n_k}\prod_{j=1}^k \theta_j^{n_j}$ with θ_j the probability that someone is born on day j, and we don't know θ_j so we might want to set an appropriate prior distribution on θ and average over it. For this, we'll need an aside about **Dirichlet integrals** on the simplex
$$\Delta_k = \left\{(\theta_1, \cdots, \theta_k) : \theta_j \geq 0, \sum_{j=1}^k \theta_j = 1\right\}.$$

Theorem 79

For any parameters $\alpha_j > 0$, setting $A = \sum_{j=1}^k \alpha_j$, we have (we only integrate over the first $(k-1)$ variables because θ_k is fixed)
$$\int_{\Delta_k} \prod_{j=1}^k \theta_j^{(\alpha_j - 1)} d\theta_1 \cdots d\theta_{k-1} = \frac{\prod_{j=1}^k \Gamma(\alpha_j)}{\Gamma(A)}.$$

For example, if all $\alpha_j = 1$, then we have the uniform distribution and this is saying that the volume of the simplex Δ_k is $\frac{\Gamma(1)^k}{\Gamma(k)} = \frac{1}{(k-1)!}$. Therefore $(k-1)!d\theta$ is the uniform probability measure on the k-simplex. We can thus use this

for a calculation: we have a uniform prior on the θs, and once we pick θ we drop n boxes into k boxes with probability $(\theta_1, \cdots, \theta_k)$. We then have

$$\begin{aligned}
\mathbb{P}(n_1, \cdots, n_k) &= \int \binom{n}{n_1, \cdots, n_k} \theta_1^{n_1} \cdots \theta_k^{n_k} (k-1)! d\theta \\
&= \binom{n}{n_1, \cdots, n_k} (k-1)! \int \theta_1^{n_1} \cdots \theta_k^{n_k} d\theta \\
&= \binom{n}{n_1, \cdots, n_k} (k-1)! \frac{\prod_i \Gamma(n_i+1)}{\Gamma(n+k)} \\
&= \frac{n!}{n_1! \cdots n_k!} (k-1)! \frac{n_1! \cdots n_k!}{(n+k-1)!} \\
&= \frac{1}{\binom{n+k-1}{k-1}}.
\end{aligned}$$

So actually a uniform randomized Maxwell-Boltzmann gives Bose-Einstein as well! As a special case (and this was Bayes' original argument in the 1780s), with $k = 2$ we're saying that

$$\int_0^1 \binom{n}{j} \theta^j (1-\theta)^{n-j} d\theta = \frac{1}{n+1}.$$

This means that if we pick a θ uniformly and then flip a θ-coin n times, we have an equal chance of getting any number out. So that's the justification for using a uniform prior – if we don't know anything, we should expect any of the outcomes to be equally likely!

Fact 80

Bayes used the following "billiard ball argument:" if we first put a red ball uniformly on $[0, 1]$ and then put n white balls uniformly on $[0, 1]$ independently as well, then the number of white balls to the left of the red ball is binomial with parameters (n, θ). But since all $n + 1$ balls are identical, the probability of the red ball being in any rank is equal.

And notice that "by thinking alone" this gives us the beta integral as well, just by dividing both sides of the above equation by $\binom{n}{j}$. And then both sides are analytic in n and j and agree on integers (and we also have Carlson's theorem), so they agree everywhere:

$$\int \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The argument for proving the general Dirichlet integral in Theorem 79 is exactly the same: put $k - 1$ red balls down uniformly in the unit interval and then put n white balls down; the number of balls in each red bucket is exactly giving us the Bose-Einstein counts in our k boxes.

Remark 81. *There's also a similar construction called the Selberg integral, but there is no known probabilistic interpretation of it yet. And we can check Professor Diaconis' paper "Five Stories for Richard" for more.*

We'll now return to a fourth appearance of the Bose-Einstein distribution:

- In **Polya's urn**, we put k balls labeled with colors $1, 2, \cdots, k$ into an urn. We then repeat the following process: take a ball in the urn at random, write down its color, and replace it and add another ball of the same color. This generates a color process $X_1, X_2, \cdots$.

Theorem 82

In Polya's urn, we have

$$\mathbb{P}(\text{within first } n \text{ draws we get } n_j \text{ of color } j \text{ for all } j) = \frac{1}{\binom{n+k-1}{k-1}}.$$

So in particular this gives us an easy way to sample from Bose-Einstein, and so does the Bayesian interpretation. We'll now do some applications with what we've done so far:

Example 83

Returning to the Bayesian birthday problem now, if we put a uniform prior on $\theta_1, \cdots, \theta_{365}$, then it turns out we only need 16 people instead of 23 to get favorable chances of an overlap.

Basically, we're saying that we want the probability that a Bose-Einstein configuration has no two balls in the same box, which is (using the Polya's urn interpretation)

$$\mathbb{P}(\text{all distinct}) = \left(\frac{k-1}{k+1}\right)\left(\frac{k-2}{k+2}\right)\cdots\left(\frac{k-(n-1)}{k+(n-1)}\right).$$

Again putting everything in the exponent and doing leading-order expansions, we can get asymptotics:

Theorem 84

With $n, k \to \infty$ and $\frac{n^2}{k} \to \lambda$, we have

$$\mathbb{P}(\text{birthday match}) \sim 1 - e^{-\lambda}.$$

So with $n = 0.83\sqrt{k}$ we get this to be $\frac{1}{2}$.

Example 85

Similarly we can also do a **Bayesian coupon collector's problem** as well: classically (due to Laplace), we uniformly drop n balls into k boxes and want at least one ball in every box. We have by inclusion-exclusion

$$\mathbb{P}(\text{all covered}) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \left(1 - \frac{j}{k}\right)^n,$$

but that can be a bit difficult to analyze. It turns out that if $n = k \log k + \theta k$ for some constant θ, then the probability of covering all boxes is asymptotic to $e^{-e^{-\theta}}$.

Feller described this in the following way: in a village of 2300 people, there is a 50-50 chance that someone is born on each day of the year. And we can do many Bayesian variations on this, but the easiest one to study is the Bose-Einstein case:

Theorem 86

Assuming a uniform prior on θ (that is, under Bose-Einstein), the chance of at least one person having each birthday is

$$\mathbb{P}(\text{cover}) = \frac{\binom{n-1}{k-1}}{\binom{n+k-1}{k-1}}.$$

(Indeed, we have to assign one ball to each box, and then the remaining balls can go in any configuration.) And the asymptotics for this are pretty different as well:

Theorem 87

Under Bose-Einstein allocation, if $n, k \to \infty$ with $\frac{n}{k^2} \to \theta$, then $\mathbb{P}(\text{cover}) = e^{-1/\theta}\left(1 + O(\frac{1}{n})\right)$.

So for $k = 365$, we now need a village of size $n = 191844$ instead! What we've done in this lecture is understand some features of Bose-Einstein statistics, and for much more on this topic, we can see Professor Diaconis' paper with Holmes "A Bayesian peek into Feller volume 1" or the textbook "Urn models and their application: An approach to modern discrete probability theory" by Johnson and Kotz.

13 April 28, 2025

We'll finish off Bose-Einstein today with a few additional remarks. As usual, we have S_n acting on C_k^n, with orbits indexed by nonnegative integers $(n_1, \cdots, n_k)$ with $\sum_i n_i = n$, and we're interested in the induced uniform distribution over all such orbits.

Last time, we described two algorithmic ways to sample from this measure. First of all, we can take the "Bayesian approach" of picking $(\theta_1, \cdots, \theta_k)$ from the k-simplex and then dropping n balls into k boxes independently with probabilities $\theta_1, \cdots, \theta_k$. (The first step here is easy, since we can let $\theta_i = \frac{X_i}{\sum_{j=1}^k X_j}$ for X_i iid standard exponential random variables.) And secondly, we can use a sequential scheme via Polya's urn.

Remark 88. *One way to remember Polya's urn is that maybe we're at Niagara Falls and want to pick a restaurant; at the beginning the first person picks one at random, but then the next people who choose will bias towards restaurants with more people. Then occupancy will be dictated by Bose-Einstein.*

There are also other ideas we can consider, such as **conditional geometric random variables**. If we let W_i be iid random variables such that $\mathbb{P}(W_i = a) = (1-t)t^a$ for nonnegative integers a, we get that

$$\mathbb{P}\left(W_1 = n_1, W_2 = n_2, \cdots, W_k = n_k \middle| \sum_{i=1}^k W_i = n\right) = \frac{1}{\binom{n+k-1}{k-1}}.$$

So like we've seen before, randomizing n helps us sample easily, and we can pick t in whatever way we'd like (in particular, to choose the expected sum to be near n). And this turns out to be more generally useful – this is an example of the Boltzmann sampler, which we'll spend a week on later on.

Example 89

Suppose we want to know the distribution of $\max(n_1, \cdots, n_k)$ for $(n_1, \cdots, n_k)$ from Bose-Einstein (that is, how many balls are in the fullest box).

This is not very easy to do, but one idea is that letting $M_k = \max(n_1, \cdots, n_k)$, we have by Bayes' theorem that

$$\begin{aligned}\mathbb{P}(M_k \le m) &= \mathbb{P}\left(\max_{1\le i\le n} W_i \le m \middle| \sum_{i=1}^k W_i = n\right) \\ &= \mathbb{P}\left(\sum_{i=1}^k W_i = n \middle| \max_{1\le i\le n} W_i \le m\right) \cdot \frac{\mathbb{P}\left(\max_{1\le i\le n} W_i \le m\right)}{\mathbb{P}\left(\sum_{i=1}^k W_i = n\right)}.\end{aligned}$$

But now we've represented the quantity of interest as a function of three calculations, each of which involves independent (in fact iid) random variables. The red term is easy to understand:

$$\mathbb{P}(W_1 \le m)^k = e^{k \log \mathbb{P}(W_1 \le m)} = e^{k \log(1-e^{-m})} \sim e^{-kt^{m+1}}.$$

Similarly, the blue term can be studied by the local central limit theorem (which analyzes the probability that S_n is equal to a particular value, rather than a region of size $\sqrt{n}$), but in this particular case we know exactly what the random variable is: we have a **negative binomial distribution** with

$$\mathbb{P}\left(\sum_{i=1}^{k} W_i = n\right) = \binom{n+k-1}{k-1}(1-t)^k t^n.$$

(In general, we would instead need to approximate with some Gaussian-type thing and end up with an expression at the density level like $\frac{1}{\sqrt{2\pi k}} \exp\left(-\frac{(m-\mu_k)^2}{\sigma_k}\right)$. And this holds for integer-valued random variables with a finite mean and variance, or continuous random variables with a density satisfying the Cramér condition, but not general random variables.) From here, what's left is the remaining fraction, which we can rewrite using the "Bayes trick" as

$$\mathbb{P}\left(\sum_{i=1}^{k} W_i = n \middle| \max_{1 \le i \le n} W_i \le m\right) = \mathbb{P}\left(\sum_{i=1}^{k} Y_i = n\right)$$

with Y_i the "truncated geometrics" satisfying $\mathbb{P}(Y_i = a) = \frac{(1-t)t^a}{1-t^{m+1}}$, and then we can just use the local CLT idea we had above.

Fact 90

With this, we can answer the problem pretty well: we already know that the maximum is likely to be 1 if $n < \sqrt{k}$ by the birthday problem. For $n = k$, it turns out it's best to take $t = \frac{1}{2}$, and we claim the answer will be around $\log n$. We want to pick it so that $e^{-n(1/2)^{n+1}}$ has a limit, but because $\log_2 n$ isn't an integer we can't actually take a nice limit so that $\frac{M_n - a_n}{b_n}$ tends to something nice! Instead, we take $m = \lfloor \log_2 n \rfloor$ and then we end up with some oscillations due to the remaining fractional part.

This oscillation is an example where we have a measure

$$\nu_n(a) = \frac{1}{Z} \exp\left(-\left(\frac{1}{2}\right)^{a+\{\log_2 n\}}\right), \quad -\infty < a < \infty,$$

where as n varies, it wiggles around because of the fractional part $\{\log_2 n\}$. We're then interested in studying some other measure like

$$\mu_n(a) = \mathbb{P}\left(\max_{1 \le i \le n} W_i \le \lfloor \log n \rfloor + a\right),$$

and we have $||\mu_n - \nu_n||_{\text{TV}} \to 0$. So these measures "merge but do not converge," and Professor Diaconis has studied things of this kind in the past. We can see his paper with Aristotile and Friedman "On Merging of Probabilities" for more details.

Remark 91. *If we look at the maximum of standard normals, we know the distribution very well: it's around $\sqrt{2 \log n}$, and we know exactly in what ways the correction terms and fluctuations work (converging to an extreme value distribution). But if we round our normal random variables for example to the nearest integer, the distribution of the largest of n such rounded normals also oscillates, but instead of extreme value distribution we get one of two values in the limit, and the probability it takes the smaller of the two values oscillates as $n \to \infty$.*

All of this being said, it's perhaps not recommended for us to go and study more Bose-Einstein statistics and try to discover new facts. This is because there's a book "Combinatorics of compositions and words" by Heubach and Mansour which already answers just about any question we might care about, and it has references to various special cases.

Remark 92. *We've been doing the "pick from uniform on the simplex Δ_k and then sample multinomial with those θ_is" here, and there's a generalization called the* **Dirichlet multinomial** *distribution where we pick from D_α on Δ_k (that is, from density $\frac{\Gamma(A)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$; this for example comes up often in non-parametric Bayesian statistics and we can see the Bayesian peak paper mentioned last lecture.*

Returning now to the Burnside process, we showed that when $k = 2$, we have (here K_0^ℓ means the Markov chain started at 0 and then run for ℓ steps, and TV is total variation distance)

$$||K_0^\ell - \pi||_{\text{TV}} \leq \left(1 - \frac{1}{\pi}\right)^\ell,$$

meaning that a constant number of steps is sufficient for convergence to stationarity. We can use the same argument for general k, but we instead get that $||K_0^\ell - \pi||_{\text{TV}} \leq (1 - f(k))^\ell$ for $f(k) \sim \frac{c}{k!}$. And this is pretty bad, because it means we'll need to take $k!$ steps to make things small. We can get a much better bound using a different strategy:

Theorem 93 (Aldous)
For all k, n, let K be the Markov kernel for the Burnside process (C_k^n, S_n), and let π be its stationary distribution. Then for any $x \in C_k^n$,

$$||K_x^\ell - \pi||_{\text{TV}} \leq n\left(1 - \frac{1}{k}\right)^\ell.$$

It's nice in particular that we have an exact bound without any big-O unknown constants, and also that this works from any starting state. What this shows is that we need at most $k \log n$ steps to stationarity, which is much better of a bound for example when $n = k$. (The actual answer is still open, though; there's no matching lower bound.) We'll do the proof next time – it's nearly computation-free!

14 April 30, 2025

We'll start by proving the Aldous coupling from last time, which states that for the Burnside process on (C_k^n, S_n), we have (for K the corresponding Markov kernel)

$$||K_x^\ell - \pi||_{\text{TV}} \leq n\left(1 - \frac{1}{k}\right)^\ell.$$

In particular, this means that the total variation distance to stationarity is at most e^{-c} after $\ell = k(\log n + c)$ steps.

Proof of Theorem 93. The proof needs a lemma:

Lemma 94
Let F_1, F_2 be any two finite sets (possibly with intersection). Then we can couple uniform permutations σ^1, σ^2 of F_1 and F_2, labeling their cycles $\{C_1^j\}$ and $\{C_2^j\}$, so that

$$C_1^j \cap (F_1 \cap F_2) = C_2^j \cap (F_1 \cap F_2).$$

In other words, we can choose the cycles of our permutations so that they intersect $F_1 \cap F_2$ in the same way.

Proof of lemma. The key fact is as follows: for any finite set S and any subset $T \subseteq S$, we can let $\sigma \in S_S$ be a uniform permutation. Write it in cycle notation, and then cross out all of the elements in $S \setminus T$. (For example if $S = \{1, \cdots, 10\}$ and $T = \{1, \cdots, 5\}$, we might get the cycle decomposition $(1, 3, 5, 2, 7)(4, 6)(8, 9, 10)$, which then becomes $(1, 3, 5, 2)(4)$.) The result is then always a uniform permutation in S_T.

So now what we do is let σ be a permutation on $F_1 \cup F_2$, and then our coupling lets σ^1 be the restriction of σ to F_1 and σ^2 the restriction of σ to F_2. And the cycles restricted to $F_1 \cap F_2$ are exactly determined by σ, so the required relation indeed holds. □

Using this lemma, we can thus construct a certain **bivariate Markov coupling** which will show that our total variation distance gets sufficiently small. We'll specify a method $\vec{K}$ for taking one step

$$(X^1, X^2) \to (Y^1, Y^2)$$

(where $X^1, Y^1, X^2, Y^2 \in C_k^n$) as a Markov chain, and it will have the property that marginally we get our desired chain:

$$\vec{K}(X^1 = x^1, Y^1 = y^1) = K(x^1, y^1), \quad \vec{K}(X^2 = x^2, Y^2 = y^2) = K(x^2, y^2).$$

In words, such a coupling means that we are running two copies of our Burnside process at the same time, but there is some way in which they are related. We will define $\vec{K}$ as a two-step process:

1. For each $a \in \{1, \cdots, k\}$, define the sets

$$F^{1,a} = \{i : X_i^1 = a\}, \quad F^{2,a} = \{i : X_i^2 = a\}.$$

   These sets are the indices in $\{1, \cdots, n\}$ which currently read a in X^1 and X^2. respectively. Remember that in the Burnside process, we pick a uniform permutation from each $F^{1,a}$ to get a uniform permutation fixing X^1. To do this in a coupled manner, we construct uniform permutations on these sets $\sigma^{1,a}$ and $\sigma^{2,a}$ satisfying the property of the lemma (that is, when the cycles overlap, they agree), and then σ^1 (resp. σ^2) is the union of all permutations $\sigma^{1,a}$ (resp. $\sigma^{2,a}$).

2. Now the second part of the Burnside process uniformly assigns each cycle one of the values $\{1, \cdots, k\}$. We do this in a coupled way by picking a single uniform $\alpha_j^a \in \{1, \cdots, k\}$ for each (a, j) and defining

$$Y_i^1 = \alpha_j^a \text{ if } i \in C_j^{1,a}, \quad Y_i^2 = \alpha_j^a \text{ if } i \in C_j^{2,a}.$$

   (each i is in exactly one cycle, so this properly defines an n-tuple for each of Y^1 and Y^2). In words, this means that for the same value of a and for the same index of cycle j, we pick the **same** random $\{1, \cdots, k\}$ to assign to those matching cycles.

The point is that the number of coordinates where things match up will only grow: if $X_i^1 = X_i^2$, then $Y_i^1 = Y_i^2$, and if $X_i^1 \neq X_i^2$, then $\mathbb{P}(Y_i^1 = Y_i^2) = \frac{1}{k}$. (Indeed, if the coordinates matched up in X, then they had the same value of a and then will also fall in the same cycle labeling j. And otherwise, they are labeled independently uniformly on $\{1, \cdots, k\}$.) Thus

$$\boxed{\mathbb{P}(Y_i^1 \neq Y_i^2) = \left(1 - \frac{1}{k}\right) \mathbb{P}(X_i^1 \neq X_i^2)}.$$

The key fact about Markov chain coupling is that the distance to stationarity can be bounded by the probability that the two components of our coupling do not agree. That is, the **coupling bound** states that if we have a bivariate Markov

chain $(X^1(\ell), Y^1(\ell))_{\ell=0,1,\dots}$ such that $X^1(0) = x$ and $X^2(0) \sim \pi$, and T is the smallest ℓ such that $X^1(\ell) = X^2(\ell)$, then

$$||K_x^m - \pi||_{\text{TV}} \leq \mathbb{P}(T > m) = \mathbb{P}\left(X^1(m) \neq X^2(m)\right).$$

And this right-hand side can be bounded: repeatedly applying the boxed identity, we find that $\mathbb{P}(X_i^1(m) \neq X_i^2(m)) \leq \left(1 - \frac{1}{k}\right)^m$ (since $\mathbb{P}(X^1(0) \neq X^2(0)) \leq 1$), and so by a union bound the right-hand side is at most $n\left(1 - \frac{1}{k}\right)^m$, as desired. □

Example 95

In case we haven't seen coupling before, here's a simpler intuitive example: we can mix a deck of cards by taking the top card and putting it in at random. To understand how long it takes for this to get to uniform stationary distribution, we'll study the inverse chain, where we take a random card out and put it at the top (this has the same mixing time).

We'll couple the cards as follows: one deck starts off sorted in order and the other starts off in random order. What we do is repeatedly uniformly pick a card name at random (for example the ace of spades), pull that card out of both decks and put it on top. This matches up the aces of spades, and from that point onward those two ace of spades will always be matched up in location in the piles. So the number of matches is monotone (it can only go up), and after every card name has been chosen once, the two decks will completely agree.

Since one deck started off uniform (and shuffling keeps it uniform), such an agreement means we're now at the stationary distribution, and thus the problem of estimating TV distance reduces to the coupon collector's problem. (The **maximal coupling theorem** then says that there always is a coupling construction which gets equality $||K_x^m - \pi||_{\text{TV}} = \mathbb{P}(T > m)$ for all m, though finding such a perfect coupling is only really possible in theory so this is a "nice but useless" theorem.)

The coupling argument of Theorem 93 is from an unpublished book "Reversible Markov Chains and Random Walks on Graphs" by Aldous and Fill (which is available online and has a chapter with various examples of coupling); we can also see Professor Diaconis' book for some other examples. And usually this technique is used for upper bounds rather than lower bounds.

Remark 96. *For $k = 2$, the Aldous coupling yields $||K_0^m - \pi||_{TV} \leq \frac{n}{2^m}$ and thus we need to take $\log n$ steps to get small. But it turns out that we know*

$$\frac{1}{4}\left(\frac{1}{4}\right)^m \leq ||K_0^m - \pi||_{TV} \leq \left(\frac{1}{4}\right)^m,$$

so a constant number of steps is actually how long it takes when started from this particular state.

Thinking about lower bounds more generally, suppose we start from the all-1s state and we have $n = k$. After one step, we pick a random permutation in S_n, and the biggest cycle is of length about $0.62n$. Thus after one step in the Burnside process, about $0.62n$ of the values will still be equal. Then that cycle can get broken up in the next step, but about $0.62^2 n$ of the values will still be equal. If we're sampling from the uniform distribution on Bose-Einstein, the biggest box count should be around $\log n$, and therefore a lower bound would be something like $\frac{\log n}{\log \log n}$. And this kind of lower-bound argument is usually easier to make as long as we can find the "slowest-mixing states."

15 May 2, 2025

We'll do an elaborate example of the Burnside process today where people do really care about running it (and where more work can be done). The overall topic is **partitions of** n: for example for $n = 4$ we have the options $1^4, 1^2 2, 13, 2^2, 4$ (so five partitions in total). Letting $\mathcal{P}(n)$ be the set of all partitions of n, one quantity that mathematicians have been interested in for a long time is $p(n) = |\mathcal{P}(n)|$. Euler wrote down the generating function

$$\sum_{n=0}^{\infty} p(n) q^n = \prod_{k=1}^{\infty} (1 - q^k)^{-1}$$

This is singular at infinitely many points on the unit circle, but we can do asymptotics using the "circle method" and it turns out that

$$p(n) \asymp \frac{1}{4\pi\sqrt{3}} e^{\pi\sqrt{2n/3}};$$

much sharper forms of this are also known. So it grows slower than $n!$, but it's still rather quickly growing – we have $p(52) \approx 250000$. To learn much more about these objects, we can see George Andrews' book "The Theory of Partitions." There are things we don't know – for example is $p(n)$ equidistributed mod 2? That is, does the proportion of values $p(1), \cdots, p(x)$ which are even tend to a limit as $x \to \infty$? (And of course, the same question could be asked for mod m in general.)

In today's class, we'll instead ask the usual question we do: pick a partition $\lambda \vdash n$ uniformly; what does it look like? Bert Fristedt's "The Structure of Random Partitions of Large Integers" is a good reference for a lot of what we'll say today.

Fact 97

Write $\lambda \sim \prod i^{a_i(\lambda)}$ if λ has a_i parts of size i. Then the number of parts of size 1 satisfies

$$\mathbb{P}_n \left(\frac{\pi}{\sqrt{6n}} a_1 \le x \right) \sim 1 - e^{-x}$$

(so we have about $\sqrt{n}$ of them) and similarly for any fixed constant j,

$$\mathbb{P}_n \left(\frac{\pi}{\sqrt{6n}} j a_j \le x \right) \sim 1 - e^{-x},$$

so that we have half as many parts of size 2, a third as many parts of size 3, and so on. Furthermore, for all $j = o(n^{1/4})$, the a_js are asymptotically independent.

We might also be curious about the largest parts:

Fact 98

Let Y_1 be the size of the biggest part of λ. Then

$$\mathbb{P}_n \left(\frac{\pi}{\sqrt{6n}} Y_1 - \log \frac{\sqrt{6n}}{\pi} \le x \right) \sim e^{-e^{-x}}.$$

(And there's a description of a stick-breaking-like process that tells us about the next largest parts Y_2, Y_3 and their joint distribution, but we won't go into it here.)

So the largest part is around $\sqrt{n}$ (with standard deviation of order $\log n$), and similarly there are about $\sqrt{n}$ parts at

the end. And since we expect about $\frac{\sqrt{n}}{j}$ parts of size j, adding that up tells us that we should expect around $\sqrt{n}\log n$ total parts, and that is indeed true:

Fact 99

The number of parts $\sum a_i(\sigma)$ of a uniform partition of n satisfies

$$\sum_{i=1}^{n} a_i \sim \sqrt{6n}\frac{\log n}{2\pi}.$$

All of this is background about partitions, and one question we might ask is "whether these theorems are any good" (for example, how accurate are they when $n = 100$). So one way we might test out that question is to generate a million partitions at random and check.

Remark 100. *Some of Professor Diaconis' students who work for the CCR needed help generating random partitions. There's a standard algorithm for generating them via Euler's formula (convert the right-hand side into something about geometric random variables and then check whether the sum of those variables is n), but when $n = 10^6$ and you want to generate a few thousand partitions, it takes far too long (since the waiting time is something like $n^{3/4}$). And the Burnside process is a way of doing this more efficiently!*

Example 101

Consider the group action of $\mathfrak{X} = G$ on itself by conjugation, meaning that $t^s = s^{-1}ts$. The orbits are conjugacy classes, and the Burnside process starts with $t \in G$ and picks a uniform s such that $t = s^{-1}ts$; that is, it picks a uniform element commuting with t. (Indeed, that's what happens in both substeps of the Burnside process, so we really only need to do it once.)

We can think of this as **nearest-neighbor random walk on the commuting graph**, where the vertices are the elements of G and we have $s \sim t$ if and only if $st = ts$. In other words, letting $C_G(t)$ be the centralizer of t,

$$K(t, t') = \begin{cases} \frac{1}{|C_G(t)|} & \text{if } t't = tt', \\ 0 & \text{otherwise.} \end{cases}$$

As a special case of our Burnside process, we thus know that K is reversible and has stationary distribution uniform on conjugacy classes:

$$\pi(s) = \frac{1}{Z|K(s)|},$$

where $K(s)$ is the conjugacy class of s and Z is the number of classes. For $G = S_n$, the classes are indexed by partitions (cycle types), so running the Burnside process will give an algorithm.

The next question is then "how do we do it," and luckily here we have a nice description of the centralizer:

$$\sigma \sim \prod_{i=1}^{n} i^{a_i} \implies C_{S_n}(\sigma) = \prod_{i=1}^{n} C_i^{a_i} \rtimes S_{a_i}.$$

(Recall from earlier in the course that this is because "conjugation is relabeling, and we can cycle each cycle and permute among the cycles of a certain size.") And these elements are easy to sample: we just need to be able to pick a_i uniforms on $\{1, \cdots, i\}$ and a uniform permutation on a_i elements.

Fact 102

Professor Diaconis gave out a handout in class which included some figures from a paper with Michael Howes "Random sampling of partitions and contingency tables: Two practical examples of the Burnside process." It shows histograms of various statistics after running a variant of the Burnside process – they're pretty close after just 20 steps even for $n = 10^4$ or 10^6, though not exact.

The thought, though, is that it is nice to lump from a chain on S_n (with $n!$ states) to the partition (with something like $e^{\sqrt{n}}$ states); as with all Burnside chains, we can do this lumping to orbits and still get a Markov chain. That's crucial in the work of most real applications, and we can write down the lumped transition matrix for example when $n = 5$:

$$K = \frac{1}{120}\begin{bmatrix} 1 & 10 & 20 & 30 & 15 & 20 & 24 \\ 10 & 40 & 20 & 0 & 30 & 20 & 0 \\ 20 & 20 & 40 & 0 & 0 & 40 & 0 \\ 30 & 0 & 0 & 60 & 30 & 0 & 0 \\ 15 & 30 & 0 & 30 & 45 & 0 & 0 \\ 20 & 20 & 40 & 0 & 0 & 40 & 0 \\ 24 & 0 & 0 & 0 & 0 & 0 & 96 \end{bmatrix}.$$

(Here, the matrix rows and columns are indexed in the order $1^5, 1^3 2, 1^2 3, 14, 2^2 1, 23, 5$.) Since our Markov chain has uniform stationary distribution, it's doubly stochastic and in fact symmetric (by reversibility). Here are some points to mention:

- If we are at the identity, then everything commutes with the identity, so when we report the conjugacy class we're in the result is proportional to the size of the class. Thus the first row and column satisfy

$$K(1^5, \lambda) = K(\lambda, 1^5) = \frac{1}{Z_\lambda}, \quad Z_\lambda = \prod_{i=1}^{n} i^{a_i} a_i!.$$

- Notice that the last row has lots of zeros. Indeed, in general if we're at the partition (n) (meaning we have an n-cycle),

$$K((n), \lambda) = K(\lambda, (n)) = \begin{cases} \frac{\phi(d)}{n} & \lambda = d^{n/d} \\ 0 & \text{otherwise.} \end{cases}$$

  In other words, the only things that commute with an n-cycle are multiples of itself, and we can end up breaking an n-cycle into smaller pieces ($\frac{n}{d}$ different d-cycles). In particular if n is a prime (like it is for $n = 5$), we have a probability $\frac{1}{p}$ of going to the identity and a probability $\frac{p-1}{p}$ of staying where we are. This is already bad news for the mixing time, since it tells us that we already need worst-case n steps to get to anything else if we start at a bad state.

- If $\lambda = \lambda_1 > \cdots > \lambda_\ell$ is a partition with all distinct parts, then the lumped Burnside chain is easy to understand: for each $1 \le i \le \ell$, we independently break λ_i into $\frac{\lambda_i}{d_i}$ pieces of size $d_i | \lambda_i$ with probability $\frac{\phi(d_i)}{\lambda_i}$. Then the union of those resulting sizes is exactly the new partition λ' (so in other words, we can only get a finer partition). However, the story is more complicated if we don't have distinct parts, because then parts of the same size can merge together into a single cycle.

We won't go through all the details of the construction here, but we do want to understand how we actually run

the lumped chain in general. If $\lambda = \prod i^{a_i}$ and $\lambda' = \prod i^{a'_i}$, we do know the formula

$$K(\lambda, \lambda') = [x_1^{a'_1} \cdots x_n^{a'_n}] \prod_i Z_{C_i^{a_i} \wr S_i}(x_1, \cdots, x_n)$$

in terms of a certain coefficient involving cycle indices. Since the centralizer of a permutation is a product of $C_i^{a_i} \wr S_{a_i}$s, we just need to be able to sample from one of those at the **partition** level.

Proposition 103

Fix ℓ and $a_\ell \geq 1$, and choose $(\lambda_j)_{j=1}^{m'}$ to be the cycle type of some element of S_{a_ℓ}. Choose a uniform $(u_j)_{j=1}^{m}$ iid in $\{1, \cdots, \ell\}$, and set $d_j = \gcd(u_j, \ell)$. We then have the parts $(b_i)_{i=1}^{\ell a_\ell}$ given by

$$b_j = \sum_{j=1}^{m} d_j \cdot 1\{\lambda_j \ell | d_j\}$$

The point is that it's not always trivial to run the lumped chain, but it's often crucial to actually do!

Remark 104. *The additional "variant" of the Burnside process (called the "reflected Burnside process") that was used in the paper uses an additional step $\lambda \mapsto \lambda^T$, which flips the partition around the diagonal line (Turning an n-cycle into an identity). So if we repeat the process where we take the transpose and apply the Burnside process in alternating order, that speeds things up significantly!*

16 May 5, 2025

The topic for today is "hit-and-run as a unifying device," named after a paper that Professor Diaconis wrote with Hans Anderson. One reason to care about the Burnside process is that it's a special case of data augmentation, auxiliary variables, and the Swendsen-Wang chain, that can all be unified into a single concept. The hope was that analysis can be done on the Burnside process and then extended to these algorithms, but it's turned out to be quite hard even in the special case.

Definition 105 (Hit-and-run for $\mathbb{R}^d$)

Let $f(x) \geq 0$ be a probability density on $\mathbb{R}^d$, possibly unnormalized (since often we cannot compute the normalizing constant easily). Our goal is to sample from f, and we do so in the following way. From $x \in \mathbb{R}^d$, pick a uniformly random point z on the unit sphere centered at x. Restrict f to the line $\ell_{x,z}$ passing through x and z and sample from the conditional distribution; the result y is one step of the Markov chain from x.

This reduces sampling in d dimensions to sampling in 1 dimension, which is a much easier problem. The claim is that this $K(x, y)$ is actually reversible and has stationary distribution given exactly by f. (The idea of the name is that we "hit a point z" and then "run to a point y" along the line.)

Remark 106. *This idea was invented by Turcin in 1971 and then rediscovered by various others. The original motivation was to sample uniformly from a compact convex set – any line intersects a convex set along a line segment, so then we just need to sample uniformly on that line. (Note however that we aren't uniform after just one step – for example if we have a long rectangle, then we're much more likely to stay around in the corners rather than moving across the rectangle.)*

If we wanted to actually do the sampling process for f restricted to a line, we could just discretize: break the line $\ell_{x,z}$ into small pieces of length ε, evaluate the value of f on each piece, and then add up those values as the normalizing constant. And if we wanted to do it better, we can shift the lattice uniformly or pick uniformly in each piece, but the point is that it shouldn't bother us too much.

Of course, there are various questions we can ask here: why do we sample z from the unit sphere, why do we do so uniformly, and why do we use f restricted to the line $\ell_{x,z}$? And if we're on some other space (a discrete problem or an infinite-dimensional space), how does this generalize? Finally, "does this actually work in practice" on high-dimensional, complicated settings, and can we prove anything about rates of convergence? It's nice that unlike something like gradient descent, we don't need anything about being unimodal or positive – the geometry doesn't prevent our algorithm from working.

People are able to prove things in special cases of hit-and-run, but they often don't talk to each other and it would be worth trying to bring the ideas across literatures. Here's a more general framework in the discrete setting which doesn't need the Euclidean structure:

Definition 107 (Hit-and-run for discrete spaces)

Let $\mathfrak{X}$ be a finite or countable set, and let $\pi(x) \geq 0$ be a probability measure perhaps only given up to a normalizing constant. To define hit-and-run, we need the following:

(a) a set of lines $\{L_i\}_{i \in I}$ for some finite or countable index set I, where each L_i is a subset of $\mathfrak{X}$. We write $I(x) = \{i \in I : x \in L_i\}$ for the set of lines passing through i.

(b) for each $x \in \mathfrak{X}_i$, a probability measure $w_x(i)$ on $I(x)$ (this used to be uniform distribution, but now it can be anything). We assume that for each fixed x, we have $w_x(i) > 0$ for some i.

(c) for each $i \in I$, a Markov kernel $K_i(x, y)$ on L_i with the specific stationary distribution $\frac{w_x(i)\pi(x)}{Z_i}$.

The **hit-and-run chain** is then the composite chain

$$K(x, y) = \sum_{i \in I} w_x(i) K_i(x, y).$$

The point is that (b) tells us how to choose lines from a point for the "hit" part, and then (c) weights the stationary distribution on the "run" part accordingly. The expression for $K(x, y)$ is then a combination over all possibilities of lines y.

Proposition 108

With the notation above, the hit-and-run chain has stationary distribution π.

Proof. We check reversibility: for any x, y, we have

$$\begin{aligned}
\sum_X \pi(x) K(x, y) &= \sum_X \sum_i \pi(x) w_x(i) K_i(x, y) \\
&= \sum_i \sum_X \pi(x) w_x(i) K_i(x, y) \\
&= \sum_i \pi(y) w_y(i) \\
&= \pi(y)
\end{aligned}$$

by Tonelli's theorem in the second line and then the definition of the stationary distribution for the K_i chains. Thus stationarity follows from reversibility of the given π. □

Notice that for any X, we can always choose $K_i(x,y) = \frac{1}{Z}\pi(y)w_y(i)$ (so we sample from the stationary distribution on the line rather than running a Markov chain with that stationary distribution); this is exactly what we were doing with the hit-and-run in $\mathbb{R}^d$ before. And also note that often $|\{i : x \in L_i\}|$ is some constant k (for example on a lattice where the lines are actual lines parallel to one of the axes); then $w_x(i) = \frac{1}{k}$ is a natural choice.

Note that if each $K_i(x,y)$ is reversible, meaning $w_x(i)\pi(x)$ is constant, then K is also reversible. (And we have to check that irreducibility and aperiodicity hold, but that's often easy in practice.) We can also extend this to general abstract measure spaces, and that's written down in a section of the paper, but we won't do that here because it won't be too helpful for the exposition

Example 109

The Burnside process is a special case of all of this. Indeed, suppose $\mathfrak{X}$ is a finite set and G is a finite group acting on $\mathfrak{X}$. Then we want to sample from the distribution $\pi(x) = \frac{1}{z}\frac{1}{|\mathcal{O}_x|}$.

What we can do is let our index set be G, the group elements, and we can define the lines

$$L_s = \langle x : x^s = x \rangle.$$

So indeed, from x we pick something from the lines containing it – that is, the set of group elements G_x that fix x – and then we choose something from L_s, and this is exactly how we described the Burnside process. But in fact notice that we can pick from any distribution on the lines as long as we adjust our Markov chain on L_s accordingly. In our case, we see that by orbit-stabilizer,

$$\pi(x) = \frac{1}{z|\mathcal{O}_x|} = \frac{|G_x|}{z|G|} \implies \pi(x)w_s(y) = \frac{1}{z|G|} \text{ constant.}$$

We'll now see a related framework which ends up being very similar to what we've already discussed:

Definition 110 (Auxiliary variables)

Let $\mathfrak{X}$ be a finite or countable set, and let $\pi(x) > 0$ be a probability distribution we want to sample from. We need the following:

(a) a set of auxiliary variables I,

(b) for each $x \in \mathfrak{X}$, a probability measure $w_x(i) > 0$ on I.

Together, these specify a joint distribution $f(x,i) = \pi(x)w_x(i)$ on $\mathfrak{X} \times I$ (pick x, then pick i conditional on x), and thus we also have the conditional distribution $f(x|i) = \frac{f(x,i)}{Z}$ for $Z = \sum_y f(y,i)$. Also, we need:

(c) for each $i \in I$, a Markov kernel $K_i(x,y)$ with stationary distribution $f(x|i)$.

The **auxiliary variables chain** is then the composite chain $K(x,y) = \sum_{i \in I} w_x(I)K_i(x,y)$.

This is all exactly the same as what we had before, except that now I is a completely abstract set of variables rather than being some set of subsets or lines:

Proposition 111

With the notation above, the auxiliary variables chain has stationary distribution π.

The point is that there is a literature of a few dozen papers on each of the constructions, but they don't seem to know much about each other! And they really are equivalent: given the setup of auxiliary variables we can set $L_i = \{x : w_x(i) > 0\}$ and get hit-and-run.

Fact 112

The picture to have in mind here is a bipartite graph with $\mathfrak{X}$ on one side and I on the other, where (x, i) is an edge with weight $w_x(i)\pi(x)$. We can then do weighted simple random walk on this bipartite graph, and if we keep track of every other step we get a chain on $\mathfrak{X}$ with the stationary distribution proportional to the weights adjacent to x. Since $\sum_i w_x(i)\pi(x) = \pi(x)$, this does have the right stationary distribution.

Example 113

Let $\mathfrak{X}$ be a finite set, and let $T_i(x) : \mathfrak{X} \to \mathbb{R}$ be a set of features for $1 \le i \le k$. Fixing $\beta_i \in \mathbb{R}$, we can then define the **exponential family**

$$p_\beta(x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{k} \beta_i T_i(x)\right).$$

We do this all the time – we have parameters for the important things in our model and those give the weights of our probability distribution. We then want to sample from p_β, and we'll do so via auxiliary variables:

(a) The set of indices is $I = [0, \infty)^k$.

(b) For each x, let $w_x(\vec{i})$ be the uniform distribution on the set

$$\{i : i_j \le \exp(\beta_j T_j(x)) \text{ for all } j\}.$$

(This is easy to sample from, since we just choose each i_j uniform from some interval.) Thus, $f(x, i)$ is uniform on the set of pairs

$$\{(x, i) : i_j \le \exp(\beta_j T_j(x))\},$$

since the exponential factors cancel out in $w_x(i)$ and $\pi(x)$.

(c) Unfortunately, going from i to y means we need to sample from the uniform distribution on points where $T_j(y)$ satisfies some condition, and depending on the problem this part can be difficult. The discovery of Swendsen and Wang is that for settings like the Ising and Potts model, this is actually doable, and we'll do that next time!

17 May 7, 2025

We've been going through some important algorithms in scientific computing – they're closely connected to the Burnside process and special cases of hit-and-run, and we'll spend today doing them a bit more. Our setting last time was **auxiliary variables** – we considered an exponential family dictated by some features $T_i : \mathfrak{X} \to \mathbb{R}$, where the distribution $\pi(x)$ is proportional to $\exp\left(\sum_j \beta_j T_j(X)\right)$. And at the end of last lecture, we described the two-step process via auxilary variables: from x, first independently pick i_j uniform on $[0, \exp(\beta_j T_j(x)]$, and then (this is the step that's hard in general) pick y uniformly among all choices satisfying $i_j \le \exp(\beta_j T_j(y))$.

Example 114

Let $\mathfrak{X} = S_n$. The **Mallows model through ℓ_2** is a probability measure on permutations

$$\pi(\sigma) = Z^{-1} \exp\left(-\beta \sum_{j=1}^{n} (\sigma(j) - \sigma_0(j))^2\right),$$

where $\beta > 0$ is some constant and σ_0 is a fixed permutation.

When $\beta = 0$ this is the uniform distribution on S_n, and as $\beta \to \infty$ this concentrates on just the single permutation σ_0 and falls off around it. These models were introduced around 1910 by psychologists who do perception experiments (in which we show seven shades of red and want participants to rank them in brightness); there σ_0 is the true ranking and there is some variability in how we would actually perform. (We also use this to measure election data where we want to understand the "mean ranking.")

Our goal will be to sample from such a distribution via a Markov chain, even though we don't know the normalizing constant Z. Without loss of generality we can let σ_0 be the identity permutation and expand out the square; the terms $\sum \sigma(j)^2$ and $\sum \sigma_0(j)^2$ are just overall constants, so we instead have

$$\pi(\sigma) \propto \exp\left(\beta \sum_{j=1}^{n} j\sigma(j)\right).$$

This is indeed of the exponential family form with all $\beta_j = \beta$, so here's what we want our auxiliary variables chain to do:

- Let $T_j(\sigma) = j\sigma(j)$. First, given σ, we want to pick $i_1, i_2, \cdots, i_n$ independently and uniformly from the intervals $0 \le i_j \le \exp(\beta j \sigma(j))$. This is easy to do.
- Then, given the values of i_j, we want to pick a permutation σ uniformly among all choices that satisfy

$$i_j \le \exp(\beta j \sigma(j)) \iff \sigma(j) \ge \frac{\log i_j}{\beta j}.$$

  Let the right-hand side be b_j; we thus want a uniform permutation τ such that $\tau(1) \ge b_1, \tau(2) \ge b_2, \cdots, \tau(n) \ge b_n$. Luckily, this is also easy to do in this case: we have n places and we want to put $1, \cdots, n$ in them, so we first look at the set of j such that $b_j \le 1$. (There's always at least one of them, because there was some permutation that fit into this recipe, specifically σ.) We let $\tau(j) = 1$ uniformly among those options. Then, we look at the set of all remaining j such that $b_j \le 2$ and let $\tau(j) = 2$ uniformly among those options. Repeat this until τ is completely filled in; we won't get stuck.

(It's worth trying to do this for a small case, like $n = 3$ – we'll learn something.)

Example 115

Let $b_1, \cdots, b_n$ be positive integers; without loss of generality suppose $b_1 \le b_2 \le \cdots \le b_n$. Then let S_b be the set $\{\sigma \in S_n : \sigma(i) \ge b_i \text{ for all } i\}$; for example if all b_i are 1 this set is all of S_n, and if $b_i = i$ then this set is only the identity permutation.

For more complicated examples, if $b_1 = \cdots = b_a = 1$ and $b_{a+1} + \cdots + b_n = 2$ then this is the set of all permutations where 1 is in one of the first a places, and if $b = (1, 1, 2, 3, \cdots, n-1)$, then the total number of possible permutations is 2^{n-1} (since there's two spots for 1 to go, then two remaining spots for 2, and so on). These sets are natural to

consider – they're called **Ferrers permutations**, and we can say a lot about them. First of all, the sets are nonempty if and only if $b_i \le i$ (because we need to be able to put the i smallest numbers somewhere) and $|S_b| = \prod_{i=1}^n (1+(i-b_i))$. Furthermore, the generating function for the inversions (minimum number of pairwise transpositions needed to invert σ) is

$$\sum_{x \in S_b} x^{I(\sigma)} = \prod_{i=1}^n (1 + x + \cdots + x^{i-b_i}),$$

and similarly the generating function for cycles is

$$\sum_{x \in S_b} x^{C(\sigma)} = \prod_{i=1}^n (x + (i - b_i)).$$

For more, we can see Diaconis, Graham and Holmes' paper "Statistical problems involving permutations with restricted positions." (And the motivation for this came from astrophysics – they wanted to see if measured redshift was correlated with something else, even though some of the data was censored.) It would be interesting to get the cycle index for this class of permutations, or find a way to study the number of fixed points.

Fact 116

If we divide the expression for either generating function by $|S_b|$, then we get generating functions for independent random variables and thus there is a probabilistic interpretation. For example,

$$\mathbb{E}_{S_b}\left[x^{C(\sigma)}\right] = \prod_{i=1}^n \left(\frac{x}{1+(i-b_i)} + \frac{i-b_i}{1+i-b_i}\right).$$

Each factor on the right-hand side is a $0-1$ valued random variable, so this tells us that $C(\sigma)$ has the same distribution as $\sum Y_i$, where Y_i are independent and Bernoulli with parameter $\frac{1}{1+i-b_i}$. Thus we have mean, variance, and the central limit theorem. Similarly, we get the sum of discrete uniforms for the number of inversions.

So something's going on with the S_bs; it's possible the cycle index polynomial is nice, but that hasn't been explored much.

Remark 117. *In this example above, we used that we can cancel out some terms by expanding out a square. Michael Howes and Chenyang Zhong managed to show that something similar (the same idea) works for any ℓ^p. This gives us a representation of a random variable, and this representation can actually be used to prove theorems such as the distribution of fixed points. (But this only really works for something like $\beta = \frac{c}{n^2}$.) This is part of the program that Professor Diaconis calls "from algorithm to theorem," and we can find some talks online of this form.*

We'll now move on to another example of auxiliary variables, but this is "the main one:"

Example 118

In **Swendsen-Wang dynamics** for the Potts model, let G be a simple connected graph (say the lattice) and let $q \ge 2$ (the number of colors) be fixed. Let $\mathfrak{X}$ be the set of all q-colorings $f : V(G) \to \{1, 2, \cdots, q\}$ of the vertices; the probability distribution we want to sample from is

$$\pi(x) = Z^{-1} \exp\left(\beta \sum_{e \in G} T_e(x)\right), \quad T_e(x) = 1\{\text{endpoints of } e \text{ are the same color}\}.$$

So T_e rewards us for being next to something else of the same color (by a factor of e^β in the probability) for each edge.

There are generalizations of this measure as well, but it'll be enough to understand just this example. To run auxiliary variables on this (also called Swendsen-Wang), we start from some configuration x and now pick auxiliary variables from the space $(0,\infty)^{|E(G)|}$ – specifically, we label edges independently with i_e uniform on $[0, e^{\beta T_e(x)}]$. Since each T_e is either 0 or 1, that means we're either uniformly on $[0,1]$ or uniformly on $[0, e^{\beta}]$.

Now for the backwards step, we must choose a uniform configuration y such that $e^{\beta T_e(y)} \geq i_e$ for all e. So if $i_e > 1$, this requires the two endpoints of e to be the same, and otherwise there is no constraint. This is again not too difficult to do once we unpack the definitions: this means that for each edge in x, if it started off bicolored then there is no constraint, and if it started off monocolored then there is a probability $\frac{e^{\beta}-1}{e^{\beta}}$ that it must be monocolored (that's the probability that i_e ended up being at least 1). Thus, the procedure is the following:

- Erase edges with different colors, and also independently erase edges with the same color with probability $\frac{1}{e^{\beta}}$.
- This gives us a new graph on $V(G)$ which can be broken up into connected components. Independently for each component, we choose that entire component to have one of the q uniform colors. This specifies the coloring for y.

The point is that this doesn't get stuck or experience "critical slowing down" – it allows us to actually run dynamics on large graphs! And so the hope is that some other examples also share that property (carrying this over to any new setting would be exciting). Unfortunately, this is "non-physical" in the sense that we usually like to use local dynamics to understand how evolution occurs under actual physical circumstances, and this algorithm is extremely non-local (big blocks are changing). But if we want to simulate from the Potts model, it's still pretty good.

For yet another reference on this, we can see Higdon's "Auxiliary variable methods for Markov Chain Monte Carlo with Applications;" the idea is that we can make the energy $\sum \beta_j T_j$ and put in an external field and it'll all still be the same.

Fact 119

It was a long open question to prove anything convergence rates or mixing times for Swendsen-Wang; the first thing that was proved is that at the critical temperature for Ising in 2D, there are two stable states and it takes a long time to go from all + to all -. But at any other temperature, it turns out to mix rapidly. We can see the paper "Efficient sampling and counting algorithms for the Potts model on $\mathbb{Z}^d$ at all temperatures," and there's lots more work done that's actively being done as well.

18 May 9, 2025

We'll continue thinking about this big class of Markov chains, thinking about the third case of **data augmentation**. The main point to get out of this is that "all of these classes are widely used, but nothing has been proved in any of the cases."

Example 120

The setup is as follows: we have a set of data in statistics, but some of it is missing and we want to fill it in. For example, say we observe N data points $(Y_i, \vec{X}_i)$ for $Y_i \in \mathbb{R}$ and $\vec{X}_i \in \mathbb{R}^d$, and the idea is that $Y = f(\vec{X}) + \varepsilon$ is some unknown function of X plus some error (for example, predict some test score given some other observables). We can do this by regression or random forests or various other approaches, but quite often there are some missing coordinates in the data points $\vec{X}_i$ (for example if our participants in a study do not provide all of the data).

One option is to throw out the missing variables, but there's still a lot of information and we don't want to bias towards data which is complete (if we even have any of it at all). For a reference, see the original paper by Tanner and Wong or the survey by van Dyk and Meng, or possibly the book "Markov Chains and Their Convergence" by Jun Liu.

Example 121

Consider the following special case: we have $X_1, \cdots, X_N$ taking values in $[k]$, and the model is that $\mathbb{P}(X_i = j) = \theta_j$ for some unknown constants θ_j which we wish to estimate. However, instead of X_i, we observe

$$Y_1 = 1\{X_1 \leq 5\}, \quad Y_2 = 1\{a \leq X_2 \leq b\}, \quad \cdots$$

(that is, we only have partial information about each observation). So there is some partition $\lambda^{(i)}$ of $[k]$ for each X_i, and all we know is the block in which each observation lands.

Data augmentation was initially phrased (and is most often used) as follows: we know that $(\theta_1, \cdots, \theta_k)$ lives in the simplex Δ_{k-1}, and we let $\pi(d\vec{\theta})$ be some prior distribution, say uniform, on Δ_{k-1}. The task is then to sample from the posterior distribution $\pi(\vec{\theta}|Y_1, \cdots, Y_N)$, and we do so as follows:

1. Start at any guess, say $\vec{\theta}^{(0)} = \left(\frac{1}{k}, \cdots, \frac{1}{k}\right)$.
2. Given this starting guess, let X_i be sampled from $\mathbb{P}(X_i = 1|Y_i, \theta^{(0)})$. For example, if the first partition for the first random variable is $\{1,2,3\}, \{4,5,6\}$ and we are told that $X_1 \in \{1,2,3\}$, then
$$\mathbb{P}(X_1^{(0)} = i) = \frac{\theta_i^{(0)}}{\theta_1^{(0)} + \theta_2^{(0)} + \theta_3^{(0)}} \text{ for } i = 1,2,3.$$
3. But now we have complete data (we have realizations of the full X_is instead of just Y_is), so we can compute the posterior by Bayes' theorem: we sample $\vec{\theta}^{(1)}$ from the posterior probability $\pi(\theta|X_1^{(0)}, \cdots, X_N^{(0)})$, which is some standard computation. Call this $\theta^{(1)}$ and feed it back into the data.

Theorem 122

With the process above (and fixed Y), the law of $\vec{\theta}^{(i)}$ converges to $\pi(\theta|Y_1, \cdots, Y_N)$ as $i \to \infty$.

This is indeed a special case of auxiliary variables, where the underlying space $\mathfrak{X}$ is the simplex (it's continuous, but that's okay) and the measure we want to sample from is $\pi(\vec{\theta}|Y_1, \cdots, Y_N)$ for some fixed Y_is. The auxiliary variables here are exactly $x_1, \cdots, x_N$, where x_is are the "conditional variables" compatible with Y_i. In this instance we have the conditional multinomials

$$W_{\vec{\theta}}(\vec{x}|Y_1, \cdots, Y_N, \vec{\theta})$$

(in the language of hit-and-run these correspond to the "lines passing through $\vec{\theta}$), and then $K_x(\vec{\theta}, \vec{\theta}') = \pi(\vec{\theta}'|x_1, \cdots, x_n)$ is the Markov chain "along the line" (in this case, just sampling directly from the stationary distribution) that gets us back to a $\vec{\theta}$.

Fact 123

So all of this provides different points of view on the same idea and possibly theorems and counterexamples, and unifying all of them is useful! These are all general stories, and in fact any Markov chain can be put in this setting, but these specific constructions can be thought about in given settings.

Example 124

We're turning back to Burnside now – we're back in the setting where a group action of G on $\mathfrak{X}$ splits the space into orbits, and we want to sample uniformly from orbits or say something about enumeration or features of those orbits. We might also be curious "how the orbits fit together" – to explain that, the leading special case of interest is the **Bruhat decomposition**.

For simplicity (and because it's the case we can do best), consider $\mathfrak{X} = GL_n(\mathbb{F}_q)$, the group of invertible matrices over a finite field. (We can take $q = 2$ if we want.) The group acting is

$$G = \mathcal{B} \times \mathcal{B}, \quad \mathcal{B} = \{\text{upper-triangular matrices in } GL_n\}$$

(often $\mathcal{B}$ is called the **Borel subgroup**), where the group action is that for any $(h, k) \in \mathcal{B}$ and any $M \in GL_n(\mathbb{F}_q)$,

$$M^{h,k} = h^{-1}Mk$$

(the point of the inverse here is to make the action compatible with the group, $(M^s)^t = M^{st}$). These are exactly the operations we do in row reduction.

Theorem 125 (Bruhat decomposition)

We have

$$GL_n(\mathbb{F}_q) = \bigsqcup_{w \in S_n} \mathcal{B}w\mathcal{B},$$

where w is interpreted as a permutation matrix.

In particular, this means there are $n!$ double cosets (that is, $n!$ orbits under this group action), and we can label them nicely with something "human-describable." Thinking about the sizes of the objects at play here, we have

$$|\mathcal{B}| = (q-1)^n q^{\binom{n}{2}}$$

(the diagonal entries are nonzero and the upper entries can be anything) and

$$|GL_n(\mathbb{F}_q)| = |\mathcal{B}| \prod_{i=1}^{n-1} (1 + q + \cdots + q^i),$$

since the first row can be any nonzero vector, the next one is any nonzero vector not in its span, and so on.

Theorem 126

For any $w \in S_n$, we have

$$|\mathcal{B}w\mathcal{B}| = |B| q^{l(w)},$$

where $l(w)$ is the number of inversions of the permutation w (that is, the number of adjacent transpositions needed to bring it back to the identity, or equivalently the number of pairs of indices $i < j$ with $w(i) > w(j)$).

The expression $\prod_{i=1}^{n-1}(1+q+\cdots+q^i)$ in the size of GL_n above is indeed consistent with this theorem, since we can count the number of inversions by adding in the numbers $1, 2, \cdots, n$ in that order and seeing how many new inversions are created (inserting j yields something between 0 and $j-1$ new inversions). And this means the double cosets have quite different sizes – they range from $|\mathcal{B}|$ to $|\mathcal{B}|q^{\binom{n}{2}}$. Furthermore, this means that the pushforward of the uniform

measure on GL_n to S_n is actually the **Mallows measure through Kendall's tau** on permutations, with the "centering point" being the identity permutation. This is a growth industry (in the sense that there are many papers) – people know quite well how to study any reasonable enumerative question on permutations. But it's also an area involving asymptotics, and people tend to care about permutations for smaller values of n.

Remark 127. *We can say more about this Bruhat decomposition in a more general setting. It turns out to be true for lots of groups, such as $GL_n(\mathbb{R})$ or $GL_n(\mathbb{C})$ (in exactly the same way) or any 'reductive group with a BN-pair" (for example, the symplectic or orthogonal groups). Instead of $\mathcal{B}$ we pick a maximal solvable subgroup, and it turns out that we will have $G = \bigsqcup_{w \in W} \mathcal{B} w \mathcal{B}$ for some* **Weyl group** *W.*

Turning back to the row-reduction perspective, suppose we have a matrix M and we want to bring it into a simpler form (which is a standard thing to do). We first look in the first column, find a nonzero element, and subtract an appropriate multiple of that row from everything else. We then repeat this process in the next column (picking a nonzero entry that isn't in one of the rows we already selected), and so on. The point is that in general, what we end up with will be a single one in each row, and the places in which they show up are exactly specifying the permutation w. To see this carefully written out in much more detail, we can see the beginning of "A century of Lie theory" by Roger Howe.

There are various extensions of this as well:

- Let $\mathcal{M} = \mathrm{Mat}_{N \times n}(\mathbb{F}_q)$ be the set of all matrices (not necessarily invertible) and let $\mathcal{B}$ be the same as above (so only invertible upper-triangular matrices). We find that

$$\mathcal{M} = \bigsqcup_{w \in W^*} \mathcal{B} w \mathcal{B}$$

  for W^* the **rook monoid** (also going under the name **complete inverse semigroup**); this is the set of $n \times n$ matrices with entries in 0 or 1, where each row and each column contain **at most** one 1, rather than exactly one.

- More generally, any reductive group has such a story; the objects in question are called **reductive monoids**, and we can see Louis Solomon's "An Introduction to Reductive Monoids" for more on this. The idea is that the "completion of GL_n is all matrices," and something similar can also be said for other settings.

19 May 12, 2025

In the last few days, we've been talking about a sweeping generalization of various algorithms, and we'll do some proofs now. The topic of today is **orthogonal polynomials and the Cannings argument for convergence of Markov chains**: there's always a question of how long it takes to go from a trick to an argument, and for something like this maybe five or so is enough.

Example 128

We'll use the following problem as an example: consider our usual Burnside process with S_n acting on C_2^n, where the orbits are the "level sets" of however many ones.

As usual, we start from a binary n-tuple and choose a permutation among the coordinates with a 0 and a permutation among the coordinates with a 1. Then we break that up into cycles and uniformly assign each of those either 0 or 1

with probability $\frac{1}{2}$; that's our new binary n-tuple. The stationary distribution is then uniform over orbits, meaning that

$$\pi(x) = \frac{1}{(n+1)\binom{n}{|x|}}.$$

Therefore the stationary distribution on the "lumped chain" $\{0, 1, \cdots, n\}$ has uniform distribution, and we've already talked about the Markov kernel for it:

$$\overline{K}(0,j) = \alpha_j^n = \frac{\binom{2j}{j}\binom{2(n-j)}{n-j}}{2^{2n}}, \quad \overline{K}(i,j) = \sum_\ell \alpha_\ell^i \alpha_{j-\ell}^{n-i}.$$

We know that this chain has various symmetries: for example, $\overline{K}(i,j) = \overline{K}(n-i,j) = \overline{K}(j,i)$.

- We've already proven previously that for all n we have $||\overline{K}_0^\ell - \overline{\pi}||_{\text{TV}} \le \left(1 - \frac{1}{\pi}\right)^\ell$ using the Doeblin technique (showing that $\overline{K}(i,j) \ge \frac{1}{\pi} \cdot \frac{1}{n+1}$, so that we can lower bound by a constant times the stationary distribution). This technique is the way that general convergence theorems for hit-and-run were first proved, and they're important when we can get them. But to prove this we needed to derive the exact formula for $\overline{K}$.
- On the other hand, we've also proved that for the unlumped ("lifted") chain, we have from any starting state x that

  $$||K_x^\ell - \pi||_{\text{TV}} \le n\left(\frac{1}{2}\right)^\ell,$$

  meaning that $\log_2 n + c$ steps are needed. This was done using coupling with some clever argument involving the permutations.

Today's class is going to discuss a third approach which will show that for all n,

$$\frac{1}{4}\left(\frac{1}{4}\right)^\ell \le ||\overline{K}_0^\ell - \overline{\pi}||_{\text{TV}} \le 4\left(\frac{1}{4}\right)^\ell.$$

(This also implies the same bounds on the lifted chain if we start from the all-zeros state.) The method of proof is **explicit diagonalization**: $\overline{K}$ is a reversible Markov chain on $\{0, 1, \cdots, n\}$ with stationary distribution $\overline{\pi}(j) = \frac{1}{n+1}$, and we will show the following:

Theorem 129

$\overline{K}$ has eigenvalues

$$0 \text{ with multiplicity } \left\lceil \frac{n}{2} \right\rceil, \quad 1 \text{ with multiplicity } 1, \quad \beta_{2k} = \frac{\binom{2k}{k}^2}{2^{4k}} \text{ with multiplicity } 1.$$

(For example for $k = 1$, the largest nontrivial eigenvalue is $\frac{1}{4}$.) Furthermore, we have a nice description of the eigenfunctions: they are the **discrete Chebyshev polynomials** $T_a(j)$ for $0 \le a \le n$, which are the orthogonal polynomials for the stationary distribution on $\{0, 1, \cdots, n\}$.

(In contrast, the ordinary Chebyshev polynomials are the stationary distribution for the continuous uniform on $[-1, 1]$.) To be more precise about what "orthogonal polynomials" means, let $L^2(\pi)$ be the set of functions $\{f : \{0, 1, \cdots, n\} \to \mathbb{R}\}$ with inner product

$$\langle f_1, f_2 \rangle = \frac{1}{n+1} \sum_{j=1}^n f_1(j) f_2(j).$$

Then the polynomials $\{1, x, x^2, \cdots, x^n\}$ viewed as functions on this space can be made orthogonal via Gram-Schmidt; the result is exactly the $\{T_0, T_1, \cdots, T_n\}$ described above. There's a more explicit way to describe them as well: we

have

$$T_0(x) = 1, \quad T_1(x) = \frac{n-2x}{n}, \quad T_2(x) = \frac{6x^2 - 6nx + n(n-1)}{n(n-1)},$$

and (this kind of stuff is always written up and accessible in some relevant resource) the general recurrence is that

$$(j+1)(n-j)T_{j+1}(x) = (2j+1)(n-2x)T_j(x) - j(j+n+1)T_{j-1}(x).$$

(So we can start the recurrence with $T_{-1} = -1$ and $T_0 = 1$.) This kind of thing is a general theme: for any measure on $\mathbb{R}$, then polynomials orthogonal with respect to that measure will satisfy a three-term recurrence of this sort. We can see from this recurrence that we are choosing the normalization so that $T_a(0) = 1$ for all a, but there are other scalings as well (for example making them monic or orthonormal); with the current one we have $\langle T_a, T_b \rangle = \delta_{ab} \cdot$ (something known in terms of a).

Remark 130. *Chebyshev used these polynomials for interpolation: we're given the values at a few points and want to get a polynomial passing through them. But they might come up in physics in the study of angular momentum, and there is a Wikipedia page with lots and lots of references as well.*

The point is that knowing all of these eigenvalues and eigenvectors lets us get bounds on rates of convergence for reversible chains. For a Markov chain with kernel $K(x,y)$ which is reversible with respect to π on a finite state space $\mathfrak{X}$, let β_i be its eigenvalues and ψ_i the corresponding eigenfunctions, where ψ_i are **orthonormal** in $L^2(\pi)$ (that is, we need $||\psi_i||_2^2 = 1$). Then we have by Cauchy-Schwarz that

$$4||K_x^\ell - \pi||_{\mathrm{TV}}^2 \le \chi_x^2(\ell) = \sum_y \frac{(K^\ell(x,y) - \pi(y))^2}{\pi(y)} = \left|\left| \frac{K_x^\ell}{\pi} - 1 \right|\right|_2^2;$$

that is, we can bound total-variation distance by chi-square distance, and then we can write that distance in terms of the eigenvalues and eigenfunctions:

$$\left|\left| \frac{K_x^\ell}{\pi} - 1 \right|\right|_2^2 = \sum_{a=1}^{|\mathfrak{X}|} \beta_a^{2\ell} \psi_a^2(x) \le \frac{1}{\pi(x)} \beta_*^{2\ell},$$

where $\beta_* = \max(\beta_1, |\beta_{\mathfrak{X}_1}|$. This inequality is often sharp, but not always, and in fact to use this at its full power (without the last inequality) we need to know the values of the eigenfunctions at the starting state.

We'll do the proof now; everything here is from Professor Diaconis' paper with Chengyang Zhong called "Hahn polynomials and the Burnside process."

Remark 131. *The Hahn polynomials are a two-parameter family of polynomials where the discrete Chebyshev polynomials are the $\alpha = 1, \beta = 1$ case, and there turns out to be a hierarchy of polynomials. The "lowest in the hierarchy" are the Hermite polynomials for Gaussians, and then we have things like Krawtchouk polynomials which are for measures with limiting Gaussians, and then above that are the Hahn polynomials. The fourth level contains the Racah polynomials, and finally the fifth level contains the four-parameter Askey-Wilson polynomials, which are orthogonal polynomial for a measure with a hypergeometric form. But that's all just for one variable, and there's much more to be said for multiple variables and lots of the theory carries over there as well (for example things like Macdonald polynomials).*

Proof of Theorem 129. We haven't talked much about how to get lower bounds on rates of convergence; they're usually easier to get, and one way we can do so is by taking ψ an eigenfunction of K with eigenvalue β (here K doesn't even need to be reversible). Then

$$||K^\ell - \pi||_{\mathrm{TV}} \ge \frac{|\psi(x)||\beta|^\ell}{2||\psi||_\infty},$$

and in this example this is what gives us our lower bound by taking ψ to be the second eigenfunction with eigenvalue 1. (This is basically using that the total variation has the alternate characterization $||K^\ell - \pi||_{\text{TV}} = \frac{1}{2}\sup_{f:||f||_\infty \le 1} |K^\ell f(x) - \pi(f)|$; here if f is a nontrivial eigenfunction, then $\pi(f)$ is just zero and $K^\ell f = \beta^\ell f$.)

The upper bound is more difficult and uses what's called the Cannings method. The idea is as follows: consider a Markov chain X with kernel $K(i,j)$ on $\{0,\cdots,n\}$ and stationary distribution $\pi(j)$, and consider the function

$$\mathbb{E}[X_1 | X_0 = x].$$

If this is equal to $ax+b$ for some constants a,b, then $\mathbb{E}[X_1 - \gamma | X_0 = x] = ax + (b-\gamma)$ and we can choose γ to get an eigenvalue: we want that $-a\gamma = b - \gamma \implies \gamma = \frac{b}{1-a}$ and thus the function $\psi(x) = x - \frac{b}{1-a}$ is an eigenfunction with eigenvalue a. So that gives a linear eigenfunction.

The next step to check is whether

$$\mathbb{E}[X_1^2 | X_0 = x] = ax^2 + bx + c$$

for some constants a,b,c. If so, then we can find constants α and β so that $\psi(x) = x^2 + \alpha x + \beta$ is an eigenfunction with eigenvalue a. The point is that if we can do this for all degrees, then the eigenvalues will be polynomials, and in fact they will be orthogonal polynomials for the stationary distribution. (And this comes up in dozens of examples in biology, mostly birth-and-death chains, where we can use this to get all of the eigenvalues and eigenfunctions). The way it's phrased in general is the following:

Theorem 132 (Cannings)
Suppose we have a Markov chain with kernel K_n on $\{0,1,\cdots,n\}$, and let π_n be the stationary distribution. Suppose that for every polynomial f of degree $\ell \le n$, $K_n f = \sum K_n(x,y) f(y)$ is a polynomial in x of degree at most ℓ. (In fact we only have to check this for $f = x^\ell$ by linearity.) Then the eigenfunctions of K_n are all of the orthogonal polynomials on $\{0,1,\cdots,n\}$.

So "checking $f = x^\ell$ by linearity" means that in fact only have to check that $\mathbb{E}[X_1^\ell | X_0 = x] = \beta x^\ell +$lower order terms, which yields an eigenvalue of β. So if the base measure is something that people care about, we can look up the eigenfunctions by consulting the relevant sources for the orthogonal polynomials! The proof of our result then follows by doing a careful analysis of the transition matrix and verifying the Cannings condition. □

Remark 133. *For a reference involving continuous-state-space chains and orthogonal polynomials, we can see Professor Diaconis' paper with Khare and Saloff-Coste "Gibbs Sampling, Exponential Families and Orthogonal Polynomials" for some applications to statistics (and some discussion about what Markov chains can have certain orthogonal polynomials as stationary distribution – this is called the Lancaster problem). In general a Markov chain cannot be diagonalized, and it's a miracle when it happens, but if it happens a bunch of times in a row it's something worth trying. And what we've seen today is an example of how we might try!*

20 May 14, 2025

We discussed the Cannings argument last time, which states that if we have a reversible Markov chain with kernel K_n on $\{0,1,\cdots,n\}$ and $K_n f(x)$ is a polynomial of degree at most $\deg f$ for all f, then the eigenfunctions of K are orthogonal polynomials with respect to π. Here's the proof:

Proof of Theorem 132. Denote the orthogonal polynomials for the stationary distribution π by $\psi_\ell(x)$ for $0 \le \ell \le n$. Clearly $\psi_0 = 1$ by convention and $K\psi_0 = 1$; this is our base case. For the inductive step, suppose we know that for

all $i \leq \ell - 1$ we have $K(\psi_i) = \lambda_i \psi_i(x)$. By self-adjointness (reversibility), we know that $\langle K(\psi_\ell), g\rangle_\pi = \langle \psi_\ell, Kg\rangle_\pi$, and furthermore because $K(\psi_\ell)$ is some polynomial of degree ℓ, we can write it as

$$K(\psi_\ell) = \lambda_\ell \psi_\ell + \sum_{i=0}^{\ell-1} a_i \psi_i$$

for some λ_ℓ and some a_i (since our orthogonal polynomials of lower degree form a basis). Taking inner products of both sides with ψ_i and using that self-adjointness property on the left side, we find that for all $0 \leq i \leq \ell - 1$,

$$a_i = \frac{\lambda_i \langle \psi_i, \psi_\ell\rangle}{\langle \psi_i, \psi_i\rangle}.$$

But these constants are all zero by the orthogonality assumption, and therefore $K(\psi_\ell) = \lambda_\ell \psi_\ell$ as desired. □

In our example our stationary distribution is uniform on $\{0, 1, \cdots, n+1\}$, and the orthogonal polynomials are called the discrete Chebyshev polynomials $T_j(x)$. One important property of these polynomials is that

$$T_j(x) = -T_j(n - x) \text{ when } j \text{ is odd},$$

and this is consistent with the property $K(i, j) = K(n - i, j)$ (the chance of winding up at j doesn't depend on whether we have i zeros and $n - i$ ones, or $n - i$ zeros and i ones). Therefore, for all j odd, we find

$$KT_j(x) = KT_j(n - x) = -KT_j(x) \implies KT_j = 0,$$

meaning that the odd-degree eigenfunctions are of eigenvalue zero. To diagonalize our chain, it thus remains to show that KT_{2a} is a polynomial of degree at most $2a$ and figure out what the leading coefficient is.

We won't go through the whole proof on the board here because it's long and technical – we can see the paper by Diaconis and Zhong for all of the details – but the main idea of the proof is that the stochastic interpretation of our operator

$$\begin{aligned} KT_{2a}(x) &= \mathbb{E}\left[T_{2a}(X_1)|X_0 = x\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[T_{2a}(X_1)|X_0 = x, \sigma_1, \sigma_2\right]|X_0 = x\right] \end{aligned}$$

can be expressed in terms of a further conditioning on the "intermediate step" σ_1, σ_2. So now we want to compute this inner sum, and since X_1 is the number of ones after flipping coins in our cycle structure, we can write everything in terms of cycle indices! Specifically, if σ_1 has a_i cycles of length i and σ_2 has b_i cycles of length i, then

$$X_1 = \sum_i i(Y_i + Z_i), \quad Y_i \sim \text{Bin}\left(a_i, \frac{1}{2}\right), \quad Z_i \sim \text{Bin}\left(b_i, \frac{1}{2}\right).$$

We now want to compute a high moment of this quantity (say to the 10th power), and we do it by expanding out the whole expression. Instead of doing that messy calculation, we'll do a **closely related** one which illustrates the main examples:

Example 134

The American Statistician is a journal similar to the American Mathematical Monthly – there was an article by Casella and George "Explaining the Gibbs Sampler" in 1992, and the first example in that paper is the following. Consider the space $\{0, 1, \cdots, n\} \times [0, 1]$ and the function on this space

$$f(j, \theta) = \binom{n}{j} \theta^j (1-\theta)^{n-j}.$$

This is a probability measure in the two variables, since summing in j yields 1 for any θ and then integrating over θ still yields 1. The Gibbs sampler then samples from the distribution by running a Markov chain alternatingly on the two components.

In more detail, we start some place, say $(j_0, \theta_0) = (n, \frac{1}{2})$. Then to get to the next step, we sample j_1 from $\text{Bin}(n, \theta_0)$ (this is the conditional distribution of j given θ_0), and then we sample θ_1 from $\text{Beta}(j+1, n-j+1)$ (this is the conditional distribution given j_1). That's one step of the Gibbs sampler – we sample the first coordinate conditional on the second, and then we sample the second coordinate conditional on the first.

The question that Professor Diaconis asked when this was first published (for a final project in this class) was to take $n = 100$, say start at $(100, \frac{1}{2})$, and prove bounds of rates of convergence. That is, find some ℓ so that $||K^\ell_{100,\frac{1}{2}} - \pi||_{\text{TV}} \leq \frac{1}{100}$. There's a standard technique called "Harris recurrence" that is used to prove convergence for Gibbs sampler. Three students spent six weeks on this and could prove that $\ell \geq 10^{33}$, but if we run numerical simulations we see that after about 50 steps the histogram of values is still far from uniform, but after about 200 steps it is close to uniform.

So Professor Diaconis tried the problem himself, and orthogonal polynomials were the key tool needed! The idea is that in this kind of bivariate chain, the first coordinate is always a Markov chain, and we can study the "$\mathfrak{X}$ chain" on $\{0, 1, \cdots, n\}$. Integrating out θ from $f(j, \theta)$ yields a uniform distribution $\pi(i) = \frac{1}{n+1}$, and we claim that

$$K(x, x') = \frac{n+1}{2n+1} \frac{\binom{n}{x}\binom{n}{x'}}{\binom{2n}{x+x'}}.$$

Indeed, this is because (doing casework on θ and integrating) $K(x, x') = \int_0^1 \binom{n}{x'} \theta^{x'} (1-\theta)^{n-x'} \cdot (n+1)\binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta$ can be evaluated with explicit beta integrals, and now we can apply the Cannings argument to K:

Theorem 135 (Diaconis, Khare, Saloff-Coste)

This Markov chain $K(x, x')$ has eigenvalues

$$\beta_0 = 1, \quad \beta_j = \frac{n(n-1)\cdots(n-j+1)}{(n+2)(n+3)\cdots(n+j+1)}.$$

In particular, $\beta_1 = 1 - \frac{2}{n+2}$, so the spectral gap is of order $\frac{1}{n}$. And the eigenfunctions are again the orthogonal polynomials for the stationary distribution – that is, the discrete Chebyshev polynomials T_j. This implies that for all starting states (j_0, θ_0),

$$\frac{1}{2}\beta_1^\ell \leq ||K^\ell_{j_0,\theta_0} - \pi||_{\text{TV}} \leq \frac{\beta_1^{\ell - 1/2}}{1 - \beta_1^\ell}.$$

Typically β_1 doesn't determine the rate of convergence alone – we often need the other eigenvalues – but in this case it does. And we see that this Markov chain does not have cutoff, since cn steps are necessary and sufficient.

Remark 136. *This general framework of two-component Gibbs sampling chains is an example of auxiliary variables! Indeed, here the state space $\mathfrak{X} = \{0, 1, \cdots, n\}$ and auxiliary set $I = (0, 1)$ are the spaces of the first and second coordinates, $\pi(x) = \frac{1}{n+1}$ is the stationary distribution on $\mathfrak{X}$, and the weight of choosing auxiliary variables $w_x(\theta)$ is exactly the conditional distribution of the second coordinate given the first. This all generalizes if $f_\theta(x)$ is binomial, Poisson, negative binomial, normal, gamma, or a sixth family which is a hyperbolic distribution; these are exactly the* **exponential families** *whose* **variance is a quadratic function of the mean**. *The prior distribution $\pi(d\theta)$ should then be the conjugate priors, and they should be beta, gamma, beta, normal, gamma, and something famous, respectively. In all cases we have polynomial eigenvectors on the $\mathfrak{X}$ chain, and in fact everything works the same on the continuous-space θ chain as well.*

There are multivariate generalizations of all of this too; we can see the paper "Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions" by Khare and Zhou.

Returning to the calculations on our beta-binomial $\mathfrak{X}$ chain, we thus want to compute

$$\mathbb{E}\left[X_1^\ell | X_0 = x\right] = \mathbb{E}\left[\mathbb{E}\left[X_1^\ell | X_0, \theta\right] | X_0 = x\right],$$

but given θ we just want to know the ℓth moment of a binomial random variable, so the inner expectation is easy:

$$\mathbb{E}_\theta[X_1^\ell] = (n - \ell + 2)_\ell \theta^\ell + \sum_{j \le \ell - 1} a_j \theta^j,$$

where $(a)_j = a(a+1)\cdots(a+j-1)$ denotes the rising factorial, and we don't need to know the lower-order terms because they're of lower degree! Then for the outer expectation the probability of θ given x we just need to do a beta integral, and the result is something like

$$\mathbb{E}_x[\theta^k] = \frac{1}{(2+n)_k} x^k + \sum_{j \le k-1} b_j x^j.$$

So plugging everything back in indeed yields what we wanted.

Remark 137. *The idea for seeing when this happens in a real example is to first compute $\mathbb{E}[X_1 | X_0 = x]$ and see whether it is linear; if so then we can try higher powers and see if there's a pattern. So having small examples and looking at data is really the secret ingredient!*

21 May 16, 2025

We've been asking questions like "can we prove anything" for special cases of the Burnside process, and we've been able to do things only in very specific models. So what we'll do instead is return to counting: we have a group G acting on a set $\mathfrak{X}$ and are interested in counting or estimating the number of orbits. We know that the Burnside process should allow us to sample uniformly from an orbit at random (after it mixes), and we'll now see how that actually occurs.

Problem 138

Let S be a finite set, and we have the ability to sample $X_1, \cdots, X_N$ uniformly at random from S. We want to convert this to an estimate on $|S|$.

If this is all we know, then one thing we can do is wait for repeats, which should take roughly $\sqrt{|S|}$ trials by the

birthday problem. More precisely, the first repeat time T satisfies the tail bound

$$\mathbb{P}\left(\frac{T}{\sqrt{|S|}} \leq x\right) \sim e^{-x^2/2},$$

and so we can repeatedly calculate values of T to get an estimate. People do try to be more careful about this, and this is called the **unseen species problem** (for example, it's asked in ecology when we collect biological species data, or if we try to estimate how many words Shakespeare knew from his writing). It's probably true that there are connections from that to what we talk about today, but things are still open for research.

The problem we're going to solve is of a different flavor, where perhaps the size of a group is something like 2^{250} and thus $\sqrt{|S|}$ is not a feasible thing to wait for:

Example 139

An idea of Broder, Jerrum, and Valient is the following: to estimate $|S|$, construct a nested decreasing filtration $S = S_N \supset S_{N-1} \supset \cdots \supset S_1$ (ending in a singleton set) such that the successive ratios $\frac{|S_j|}{|S_{j+1}|}$ are not too small and also not too close to 1. Sampling from $S = S_N$ uniformly then allows us to estimate $\frac{|S_{N-1}|}{|S_N|}$ (by seeing what fraction of our samples lands in the smaller set), and similarly sampling from S_{N-1} allows us to estimate $\frac{|S_{N-2}|}{|S_{N-1}|}$, and so on. We can then estimate

$$|S_N| \approx \left(\frac{|S_{N-1}|}{|S_N|} \cdot \frac{|S_{N-2}|}{|S_{N-1}|} \cdots \frac{|S_1|}{|S_2|}\right)^{-1}.$$

It turns out that for a rapidly mixing Markov chain (which we can use for sampling), we can estimate these ratios and by large deviations estimates we get exponentially small errors with reasonable sample sizes. For example, we can do this to estimate the number of perfect matchings on a bipartite graph – counting exactly is #-P complete, but we can estimate it in something like n^4 time. (The Markov chain is fairly crude too – we take two edges in our matching and swap them if possible.) And there's a converse where being able to count also gives us a fast-mixing Markov chain. For more on this general literature (these are called **ratio estimators** in statistics), see Alistair Sinclair's book.

Example 140

A second key idea that will feature in our main topic today is **importance sampling**. Here, suppose μ is a probability measure on our finite set S and we want to approximate $I(f) = \sum_{s \in S} f(s)\mu(s)$ for some given function f, but we can't sample from μ. Instead, we have some other measure ν from which it is easy to sample.

What we do is sample iid $Y_1, \cdots, Y_N$ from ν, and then estimate

$$I(f) = \frac{1}{N}\sum_{i=1}^{n} f(Y_i) \cdot \frac{\mu(Y_i)}{\nu(Y_i)}$$

as long as we can write down the ratio $\frac{\mu}{\nu}$. (Indeed, the expectation of each term under ν of this quantity is exactly $I(f)$.) This is a whole subject – it's basically saying that we can tilt our measure to sample from tails instead of the bulk – and for a reference on this we can see the paper "The sample size required in importance sampling" by Chatterjee and Diaconis.

Example 141

What we'll focus on today is the paper "Counting the number of group orbits by marrying the Burnside process with importance sampling" by Diaconis and Zhong. The idea is that there are examples where we want to count orbits of a group G acting on $\mathfrak{X}$ but don't have nice names for the $\mathcal{O}_i$s, hence don't have nice ways to sample with ratio estimators.

We'll again need a sequence of sets (they don't have to be subsets anymore, but they will be in our example)

$$\mathfrak{X} = \mathfrak{X}_N,\ \mathfrak{X}_{N-1},\ \cdots,\ \mathfrak{X}_1$$

where this time we have groups G_i acting on each $\mathfrak{X}_i$

$$G = G_N,\ G_{N-1},\ \cdots,\ G_1 = \mathrm{id}.$$

We relate different terms sequence by defining maps ϕ_i for all $1 \leq i \leq n-1$, where ϕ_i maps $\mathfrak{X}_{i+1}$ **onto** $\mathfrak{X}_i$. What we need an **an efficient algorithm for counting the size of the fixed-point set**

$$\mathrm{Stab}_i(x) = \{s \in G_i : x^s = x\}.$$

We assume again that successive ratios are not too small or close to 1, so we can then try to estimate each term on the right-hand side of

$$k(\mathfrak{X}, G) = \prod_{i=1}^{N-1} \frac{k(\mathfrak{X}_{i+1}, G_{i+1})}{k(\mathfrak{X}_i, G_i)}$$

instead. The way we will do that is by importance sampling, and this relies on the following idea:

Proposition 142

Suppose T_{i+1} is random in $\mathfrak{X}_{n+1}$ with

$$\mathbb{P}(T_{i+1} = x) = \frac{1}{k(X_{i+1}, G_{i+1})|\mathcal{O}_{i+1}(x)|}.$$

(We can do this by running the Burnside process until the chain has mixed.) Then

$$\frac{|G_{i+1}|}{|G_i|} \cdot \mathbb{E}\left[\frac{|\mathrm{Stab}_i(\phi_i(T_{i+1}))|}{|\mathrm{Stab}_{i+1}(T_{i+1})| \cdot |\phi_i^{-1}(\phi_i(T_{i+1}))|}\right] = \frac{k(\mathfrak{X}_i, G_i)}{k(\mathfrak{X}_{i+1}, G_{i+1})}.$$

This follows from an elementary calculation, and the point is then that we can run the Burnside process, calculate estimates of the left-hand side, and then multiply them together across all i to get an estimate. Here's a real example where this might be useful:

Definition 143

The group $U_n(\mathbb{F}_q)$ is the set of all uni-upper-triangular matrices with entries in $\mathbb{F}_q$ (meaning that the entries on the diagonal are 1 and the entries below are 0, but the entries above the diagonal are arbitrary).

This is a natural object to consider – for $q = p^a$, this is the Sylow-p subgroup of $GL_n(\mathbb{F}_q)$. We have $|U_n(\mathbb{F}_q)| = q^{\binom{n}{2}}$, and we know a lot about the group (it's a central thing in modern Lie theory, modular representation theory, and so on). But it's not known how many conjugacy classes there are for the group, and correspondingly we don't really know the characters of $U_n(\mathbb{F}_q)$.

When $n = 3$, this is the Heisenberg group and we know everything (there are $q + (q^2 - 1)$ classes), but as n grows it's very difficult to say anything – the number of conjugacy classes has only been computed up to around $n = 16$ using lots of computing power. John Thompson has had a paper on his website for 20 years trying to prove a conjecture that the number of classes is a polynomial in q, but no progress has been made since – it's **Higman's PORC conjecture** that it's in fact of degree $\left\lfloor \frac{n(n+6)}{12} \right\rfloor$, and it's known that the number is between $q^{\frac{1}{12}n^2}$ and $q^{\frac{1}{4}n^2}$. So that's the state of the problem, and it's had quite a lot of papers and conjectures surrounding it!

Example 144

The Burnside process comes to the rescue here, and the specific one we'll use is the **commuting graph walk**: let the group act on itself via $t^s = s^{-1}ts$, and from each $s \in G$ pick a uniform t such that $st = ts$. This has stationary distribution uniform on each orbit – that is, on each conjugacy class.

Abstractly this makes sense, but we want to be able to do it in our specific example: given a matrix $M \in U_n(q)$, we want some uniform matrix M' such that $MM' = M'M$. The idea is that M and M' are each the identity matrix times some upper-triangular matrix M_1, M_1', and our goal is now to find a uniform M_1' such that $M_1 M_1' = M_1' M_1$ – this is now just a linear system in the entries of M_1, and so we can find a basis for the solution space by Gaussian elimination and then pick coefficients uniformly at random to get some element $\sum \varepsilon_j B_j$. (Since we do this at every step of the Burnside process, at each stage of our algorithm, we need this to be fast.)

To specify the algorithm from here, we need to find a nested decreasing sequence of sets. Our sets here will be **pattern subgroups**: the idea is to consider sets J of pairs (i, j) with $1 \le i < j \le n$, such that $(i, j), (j, k) \in J \implies (i, k) \in J$, and for any such J to consider the matrices which are only nonzero in the J entries

$$U_J = \{u \in U_n(\mathbb{F}_q) : u_{ij} = 0 \quad \forall (i, j) \notin J\}.$$

Concretely, for $n = 4$, we'll use the following sets:

$$\mathfrak{X}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathfrak{X}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathfrak{X}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathfrak{X}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & * & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathfrak{X}_5 = \begin{bmatrix} 1 & 0 & * & * \\ 0 & 1 & * & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathfrak{X}_6 = \begin{bmatrix} 1 & 0 & * & * \\ 0 & 1 & * & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathfrak{X}_7 = \begin{bmatrix} 1 & * & * & * \\ 0 & 1 & * & * \\ 0 & 0 & 1 & * \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Each of these is a subgroup (the property of being "closed" above is exactly what guarantees this), and we have a chain of subgroups of length $\binom{n}{2}$. So what happens is the following: for each stage i, run the Burnside process for some burn-in steps $B_i = 50000$ and then some sampling steps $N_i = 50000, 100000, 150000$ (the results are indistinguishable so we've likely already converged). This generates the matrices T_i needed in the proposition, and we estimate the value of the left-hand sides to get estimates for orbit ratios.

Fact 145

If we run the process described above, it turns out to let us estimate $k(U_n(\mathbb{F}_2))$ up to $n = 40$ (and similarly $k(U_n(\mathbb{F}_3))$). The estimated values agree very closely with the known values up to $n = 16$, and they turn out to agree closely with the predicted degree of the polynomial $\frac{n(n+6)}{12}$ more so than $\frac{n^2}{12}$.

For a sanity check, we can also try to do this process to estimate the number of orbits on the (C_2^n, S_n) chain = (which is exactly $n+1$). We can work everything out, and doing it for something like $n = 20$ shows that everything does work. But nothing has been proven yet and this hasn't been applied to other examples yet, so there's more to do!

22 May 19, 2025

We're starting a new subject today, the **Boltzmann sampler**. There's a French school of combinatorialists (often called the Flajolet school) who have developed a set of techniques called "symbolic computation" for writing down generating functions. It's quite close to formal language theory and automata theory, and often they can turn "magic tricks" into smooth machines. We can take a look at the book "Analytic Combinatorics" by Flajolet and Sedgewick – it writes down various generating functions and then uses complex variables (singularity analysis) to get information about the coefficients. The point is that another development of this school yields, from the generating function, a way of sampling from the measure induced by the coefficients. (And a stochastic representation of the measure can often be used as a proof technique, like coupling, for theorems.)

Definition 146

A **combinatorial class** $\mathcal{C}$ is a finite or countable set with a size function $|\cdot| : \mathcal{C} \to \mathbb{Z}_{\geq 0}$, such that the number of elements in $\mathcal{C}$ of any given size n is finite. We write $\mathcal{C}_n$ for the set of elements of size n and $c_n = |\mathcal{C}_n|$; the **generating function** for the class $\mathcal{C}$ is the formal power series $C(x) = \sum_{n=0}^{\infty} c_n x^n$.

We write $[z^n]C(z) = c_n$ for the nth degree coefficient, and we say that two classes $\mathcal{A}, \mathcal{B}$ are **isomorphic** if $a_n = b_n$ for all n (so sizes are the same, meaning there's a size-preserving bijection).

We'll combine these classes in various ways, and here are the some examples of **admissible constructions**. We must have an ***m*-ary construction** Φ takes any m classes $\mathcal{B}^{(1)}, \cdots, \mathcal{B}^{(m)}$ into another class $\mathcal{A} = \Phi(\mathcal{B}^{(1)}, \cdots, \mathcal{B}^{(j)})$, and the constraint is basically that we need to be able to tell the things of size n in $\mathcal{A}$ based on only the information of things up to size n in the $\mathcal{B}^{(i)}$s.

- The **product**
$$\mathcal{A} = \mathcal{B} \times \mathcal{C} = \{(\beta, \gamma) : \beta \in \mathcal{B}, \gamma \in \mathcal{C}\}, \text{ where } |\alpha| = |(\beta, \gamma)| = |\beta| + |\gamma|$$
is a 2-ary construction, and we know that
$$a_n = \sum_{k=0}^{n} b_k c_{n-k} \implies A(z) = B(z)C(z).$$
- If $\mathcal{B}$ and $\mathcal{C}$ are disjoint classes (so for example if we color the elements so that they're distinguishable from each other), then the **union** $\mathcal{A} = \mathcal{B} \cup \mathcal{C}$ is also a 2-ary construction, where the size of any $\omega \in \mathcal{A}$ is the restriction of the weight function to either $\mathcal{B}$ or $\mathcal{C}$. Then $a_n = b_n + c_n$, so $A(z) = B(z) + C(z)$.
- In the **sequence construction**, consider any combinatorial class $\mathcal{B}$. Then
$$\text{Seq}(\mathcal{B}) = \varepsilon \cup \mathcal{B} \cup (\mathcal{B} \times \mathcal{B}) \cup (\mathcal{B} \times \mathcal{B} \times \mathcal{B}) \cup \cdots$$
is the set of all finite-length ordered sequences of elements in $\mathcal{B}$, and where $|(\beta_1, \cdots, \beta_\ell)| = \sum_{i=1}^{\ell} |\beta_i|$. The

generating function is then given by

$$\text{Seq}(B)(z) = 1 + B(z) + B(z)^2 + B(z)^3 + \cdots = \frac{1}{1 - B(z)}.$$

Here we assume the existence of the **empty class** $\mathcal{E}$, which contains a single element ε of size 0. We then have $\mathcal{E} \times \mathcal{A} \cong \mathcal{A} \cong \mathcal{A} \times \mathcal{E}$ and $E(z) = 1$. It will also be useful to have the class $\mathcal{Z}$ which contains only a single element z of size 1; we then have $Z(z) = z$.

- The **multiset** construction is the set of all multisets (repetition allowed) of any finite size of $\mathcal{B}$

$$\text{MSet}(\mathcal{B}) = \text{Seq}(\mathcal{B})/R,$$

where R is the relation which makes

$$(\beta_1, \cdots, \beta_\ell) \overset{R}{\sim} (\beta_{\sigma(1)}, \cdots, \beta_{\sigma(\ell)}) \quad \forall \sigma \in S_\ell.$$

We'll describe the generating function soon.

- The **powerset** construction can be described as

$$\text{PSet}(\mathcal{B}) = \text{elements of Seq}(\mathcal{B}) \text{ with no repetitions.}$$

Note that combinatorial classes form a **semiring** under addition and multiplication using union and product (meaning that we have associativity and distributivity).

Example 147

Let $\mathcal{Z} = \{\cdot\}$ contain a single point. Then $\text{Seq}(\mathcal{Z})$ contains $\varepsilon, \cdot, \cdot\cdot, \cdots$, and so on, which is exactly the nonnegative integers $\mathbb{Z}_{\geq 0}$.

Example 148

Let $\mathcal{A} = \mathcal{Z} \cup (\mathcal{Z} \times \mathcal{Z})$. Then $\text{Seq}(\mathcal{A})$ combines together a sequence of single dots $\cdot$ and double dots $\cdot\cdot$, and the number of things of weight n are the total number of ways of putting together a sequence 1s and 2s to form n, which is exactly the Fibonacci numbers. (So $\text{Seq}(A)(z) = 1 + z + 2z^2 + 3z^3 + 5z^4 + \cdots$.)

By iterating these "basic constructions," we get a lot of common combinatorial classes – this is almost all of the common strategies we'll need.

Theorem 149

The basic constructions are admissible, and the generating functions are as follows. We already said that sum (union) yields $A(z) = B(z) + C(z)$, product yields $A(z) = B(z)C(z)$, and Seq yields $A(z) = \frac{1}{1-B(z)}$. We also have that if $\mathcal{A} = \text{PSet}(\mathcal{B})$, then

$$A(z) = \prod_{n=1}^{\infty} (1 + z^n)^{B_n} = \exp\left(\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} B(z^k) \right),$$

and if $\mathcal{A} = \text{MSet}(\mathcal{B})$, then

$$A(z) = \prod_{n=1}^{\infty} \frac{1}{(1 - z^n)^{B_n}} = \exp\left(\sum_{k=1}^{\infty} \frac{B(z^k)}{k} \right).$$

We'll prove these last two relations, and this should look familiar from Polya's theorem earlier in the course:

Proof. First, we consider the powerset. If $\mathcal{B}$ is finite, then we claim that

$$\mathrm{PSet}(\mathcal{B}) \cong \prod_{\beta \in \mathcal{B}} (\varepsilon + \beta).$$

Indeed, expanding this out we choose whether to choose each element or not, and then all of the εs go away without any effect on the sizes. (This is kind of like how $(1+a)(1+b)(1+c) = 1 + (a+b+c) + (ab+bc+ac) + abc$.) But the right-hand side is just a finite product, so

$$A(z) = \prod_{\beta \in \mathcal{B}} (1 + z^{|\beta|}) = \prod_{n=0}^{\infty} (1+z^n)^{B_n},$$

and therefore

$$A(z) = \exp\left(\sum_{n=0}^{\infty} B_n \log(1+z^n) \right) = \exp\left(\sum_{n=0}^{\infty} B_n \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} z^{nk} \right),$$

so that changes the order of summation yields

$$\exp\left(\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sum_{n=0}^{\infty} B_n z^{nk} \right) = \exp\left(\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} B(z^k) \right).$$

But now if $\mathcal{B}$ is infinite, we can just "pass to the limit." We'll say that more carefully: write $k[[z]]$ for the ring of formal power series over a field k. Let $f(z) = \sum_{n=0}^{\infty} f_n$ be an element of this ring, and let $\mathrm{val}(f)$ be the place of the smallest m such that $f_m \neq 0$ (setting $\mathrm{val}(0) = -\infty$). We can then define a distance $d(f,g) = 2^{-\mathrm{val}(f-g)}$ for any $f, g \in k[[z]]$; this is a metric and in fact an ultrametric. We then have $f^{(j)} \to f$ if and only if the coefficients stabilize, and under this topology $k[[z]]$ is a compact metric space.

This metric makes $\sum_{j=1}^{\infty} f^{(j)}$ potentially make sense as an infinite sum, and in fact it converges if and only if $\mathrm{val}(f^{(j)}) \to \infty$ as $j \to \infty$. In particular, if the zeroth term f_0 of the power series is zero, then $\sum_{j=0}^{\infty} f^j = \frac{1}{1-f}$ (meaning that whatever power series we get multiplies with $(1-f)$ to 1). We can then similarly take logs, exponentials, and derivatives in the usual way as long as we're careful about the constant terms; the usual rules about those computations still apply.

So in our example with the powerset $\mathcal{A} = \mathrm{PSet}(\mathcal{B})$, we can let $\mathcal{B}_{\leq m} = \sum_{j=1}^{m} \mathcal{B}_j$ be the set of all elements of size up to m and $\mathcal{A}_{\leq m} = \mathrm{PSet}(\mathcal{B}_{\leq m})$. But then $\mathcal{A}_n$ depends only on $\mathcal{B}_1, \cdots, \mathcal{B}_n$ (and not larger $\mathcal{B}_j$s), so we have

$$\mathcal{A}(z) = \mathcal{A}_{\leq m}(z) + O(z^{m+1}), \quad \mathcal{B}(z) = \mathcal{B}_{\leq m}(z) + O(z^{m+1}),$$

where the big-O notation denotes that we have no terms of z^m or smaller. Since $\mathcal{A}_{|lem}$ and $\mathcal{B}_{\leq m}$ are related by the exponential relation we claimed above, we can let $m \to \infty$ and the claim follows.

Similarly for multisets, we first let $\mathcal{B}$ be finite. Then

$$\mathrm{MSet}(\mathcal{B}) = \prod_{\beta \in \mathcal{B}} \mathrm{Seq}(\beta),$$

since we can order the finite alphabet and then any element of the multiset has some finite number of elements of each kind. Therefore

$$A(z) = \prod_{\beta \in \mathcal{B}} \frac{1}{1 - z^{|\beta|}} = \prod_{n=1}^{\infty} \frac{1}{(1-z^n)^{B_n}},$$

and again putting things up in the exponent yields

$$\exp\left(\sum_{n=1}^{\infty} -B_n \log(1-z^n)\right) = \exp\left(\sum_{k=1}^{\infty} \frac{B(z^k)}{k}\right)$$

by doing the same swapping of order of summation. Then we pass to the limit again to do it for finite $\mathcal{B}$. □

This calculus can be abstracted more into the theory of **species**, which is essentially the same thing but said in category language. We can look this up on our own and find some good books, but it's all generating functions (and thus the problems it applies to are those with nice generating functions). Those are called "exactly solvable models," and unfortunately lots of problems in the world don't have such nice generating functions. So thinking like a probabilist still has lots of value when we can't use these techniques! We'll see some applications and abstractions next time.

23 May 21, 2025

We've been discussing the "symbolic method," in which we try to compute the generating functions of various combinatorial classes $\sum_{n\geq 0} c_n z^n$. We started with the empty class and the class $\mathcal{Z}$ of a single element, and then we found that there were a set of admissible constructions that we could use to get more complicated classes (union, product, sequence, powerset, multiset). Everything we're doing today will (continue to) count **unlabeled structures**, and we'll just do examples today to properly learn how this all works. The point is that when we start using all of this with probability, we'll be familiar with the necessary tools.

Fact 150

We'll let $\mathcal{I} = \text{Seq}(\mathcal{Z}) \setminus \varepsilon = \{1, 2, \cdots\}$ denote the positive integers, so that the generating function is $I(z) = \frac{z}{1-z}$.

Example 151

We write $\lambda \vDash n$ for **compositions** of n (meaning that we divide n into nonzero parts, **and order matters**). Then $\mathcal{C} = \text{Seq}(\mathcal{I})$, so the generating function for compositions is

$$C(z) = \frac{1}{1-I(z)} = \frac{1}{1-\frac{z}{1-z}} = \frac{1-z}{1-2z}.$$

Expanding out this function, we thus find that

$$C(z) = \sum_{n=0}^{\infty} 2^n z^n - \sum_{n=0}^{\infty} 2^n z^{n+1} \implies C_n = 2^{n-1} \text{ for all } n \geq 1.$$

And this makes sense, because by a "stars-and-bars" argument we can decide at each of $(n-1)$ spots whether to start a new part in the composition or not.

Example 152

We write $\lambda \vdash n$ for **partitions** of n (this time order does not matter), so that $\mathcal{P} = \text{MSet}(\mathcal{I})$. Thus the generating function is given by

$$P(z) = \exp\left(I(z) + \frac{I(z^2)}{2} + \frac{I(z^3)}{3} \cdots\right) = \prod_{i=1}^{\infty} \frac{1}{1-z^m},$$

and this is indeed the usual generating function we have for partitions.

There is also a way of putting restrictions on our constructions:

Definition 153

Let $\tau \subseteq \mathcal{I}$ be some subset, and let κ be any of our basic constructions above. We write $\kappa_\tau(\mathcal{A})$ for the construction which applies κ to $\mathcal{A}$ but with the number of parts only in the set τ.

For example, $\mathcal{B} = \text{Seq}_{=k}(\mathcal{A})$ means that we must take exactly sequences of length k, and thus this is exactly the k-fold product of $\mathcal{A}$ with itself and $B(z) = A(z)^k$. And similarly if $\mathcal{B} = \text{Seq}_{\geq k}(\mathcal{A})$, we have $\mathcal{B} = \mathcal{A}^k \times \text{Seq}(\mathcal{A})$ and thus

$$B(z) = \frac{A(z)^k}{1 - A(z)}.$$

We'll use this construction in our next example:

Example 154

Let τ be a subset of the positive integers, and let

$$\mathcal{C}^\tau = \text{Seq}(\text{Seq}_\tau(\mathcal{Z})), \quad \mathcal{P}^\tau = \text{MSet}(\text{Seq}_\tau(\mathcal{Z})).$$

Unpacking the definition, $\text{Seq}_\tau(\mathcal{Z}))$ is the set of all integers in τ, and thus $\mathcal{C}^\tau$ is the set of all compositions with parts in τ and $\mathcal{P}^\tau$ is the set of all partitions with parts in τ. We can then compute the generating functions because

$$\mathcal{A} = \text{Seq}_\tau(\mathcal{Z}) \implies A(z) = \sum_{n \in \tau} z^n,$$

so plugging into the usual construction yields

$$\mathcal{C}^\tau(z) = \frac{1}{1 - A(z)} = \frac{1}{1 - \sum_{n\in\tau} z^n}, \quad \mathcal{P}^\tau(z) = \prod_{n \in \tau} \frac{1}{1 - z^n}$$

and these are indeed quantities that we might encounter in analytic number theory. For example if $\tau = \{1, \cdots, r\}$ we can explicitly compute and find that

$$\mathcal{C}^{\{1,\cdots,r\}}(z) = \frac{1}{1 - z - z^2 - \cdots - z^r} = \frac{1 - z}{1 - 2z + z^{r+1}};$$

this yields the **generalized Fibonacci numbers** and we can get formulas for asymptotics if we can solve for the roots of the corresponding polynomial in the denominator.

Similarly, we can always ask the analogous questions for partitions – for example, "how many ways can we make change with a dollar using pennies, nickels, dimes and quarters?" comes out of the coefficient of the generating function

$$[z^{100}] \frac{1}{(1 - z)(1 - z^5)(1 - z^{10})(1 - z^{25})} = 213.$$

Example 155

The number of compositions with exactly k parts is given by

$$\mathcal{C}^{(k)} = \text{Seq}_k(\mathcal{I}) = \mathcal{I} \times \cdots \times \mathcal{I},$$

so $C^{(k)}(x) = \left(\frac{z}{1-z}\right)^k$.

Expanding this out yields the usual formula for stars-and-bars (shifted accordingly).

Example 156

The number of partitions with at most k parts is

$$\mathcal{P}^{(\leq k)} = \mathrm{MSet}_{\leq k}(\mathcal{I}) \implies P^{(\leq k)}(z) = \prod_{m=1}^{k} \frac{1}{1-z^m}.$$

With this last example, we come to an interesting problem: we know that we can represent partitions with dot diagrams, and we say the **Durfee square** is the largest $h \times h$ block of dots that we can fit inside the partition. Notice that we can always break up a partition into its Durfee square, a partition to the right of it with at most h parts, and a partition below it with all parts of size at most h. Therefore this decomposition tells us that as combinatorial classes,

$$\mathcal{P} = \bigcup_{h=1}^{\infty} \mathcal{Z}^{h^2} \times \mathcal{P}^{\leq h} \times \mathcal{P}^{\{1,2,\cdots,h\}},$$

so in particular the partition generating function satisfies (by flipping over the diagonal the classes $\mathcal{P}^{\leq k}$ and $\mathcal{P}^{\{1,2,\cdots,k\}}$ are isomorphic)

$$\prod_{n=1}^{\infty} \frac{1}{1-z^n} = \sum_{k=1}^{\infty} \frac{z^{h^2}}{(1-z)(1-z^2)\cdots(1-z^h))^2}.$$

Fact 157

This h is actually the h-index used in academia – it's the maximum value such that someone has written h papers that have been cited h times. Seiberg (the physicist) pointed out that the total number of citations that someone receives is roughly $4h^2$ by empirical observation, but there's a story to be told here. For example, there's many possible distributions on partitions, and we know the limit shape for most nice ones – it would be interesting to see what we can get theoretically out of each model and which one fits the citation distribution best.

It turns out that lots of partition identities end up being nice in Frobenius coordinates (the "remaining parts" other than the Durfee square), and we can see Macdonald's book for more in that direction.

Example 158

A **rooted plane tree** is a tree whose vertices are unlabeled, but where we can't switch the order of the children of any vertex. This is an "important recursive definition in computer science," and we can enumerate by removing the root.

Any binary plane tree is some (possibly empty) sequence of rooted plane trees attached to the root, and therefore if $\mathcal{G}$ is the set of rooted plane trees,

$$\mathcal{G} = \mathcal{Z} \times \mathrm{Seq}(\mathcal{G}) \implies G(z) = \frac{z}{1-G(z)}.$$

This is a quadratic equation in G, so

$$G(z)^2 - G(z) + z = 0 \implies G(z) = \frac{1 \pm \sqrt{1-4z}}{2},$$

and we take the negative root so that it makes sense at $z = 0$. We then end up with

$$G(z) = \sum_{n \geq 1} \frac{1}{n} \binom{2n-2}{n-1} z^n,$$

and thus $G_n = C_{n-1}$ is a Catalan number.

Example 159

A **Polya tree** is a rooted tree whose vertices are again unlabeled, but where we can now switch the order of children.

Letting $\mathcal{PT}$ denote the set of Polya trees, we now get

$$\mathcal{PT} = \mathcal{Z} \times \text{MSet}(\mathcal{PT}) \implies \text{PT}(z) = z \exp\left(\sum_{j=1}^{\infty} \frac{\text{PT}(z^j)}{j}\right).$$

This is more difficult to work with than what we had before, but it can be done – we can get sharp asymptotics via **Otter's method**, and additionally we can use the Boltzmann sampler to get an algorithm for generating random trees on a computer to study whatever feature we want.

24 May 23, 2025

Today, we'll talk about formal language theory in a way that relates the concepts to probability. (All of this comes from a book Flajolet and Sedgewick, as well as a paper by Mireille Bousquet-Mélou.) We've been doing things with combinatorial classes and various funny operations, and this will have its own language:

Definition 160

Let $\mathcal{A}$ be a finite alphabet, where each letter (also called "atom" or "element") in $\mathcal{A}$ has **weight** 1. (Thus, the generating function here is $A(z) = mz$, where $m = |\mathcal{A}|$.) Let $\mathcal{W}$ be the set of all words in $\mathcal{A}$ (this is sometimes also denoted $\mathcal{A}^*$); a word is assigned a weight which is the sum of the weights of the letters (that is, the length), and so $\mathcal{W} = \text{Seq}(\mathcal{A}) \implies W(z) = \frac{1}{1-mz}$ and $W_n = m^n$.

Definition 161

With the notation above, a **language** is any subset of $\mathcal{W}$. A language $\mathcal{L}$ is **regular** if there is some combinatorial class $\mathcal{M}$ built from a finite alphabet and finite iterations of the operations union, product, and Seq, where $\mathcal{M}$ is isomorphic to $\mathcal{L}$ (that is, $L_n = M_n$ for all n).

Example 162

Suppose our alphabet has two letters $\mathcal{A} = \{a, b\}$. Each word w can be decomposed by appearances of b (for example, `aaabaababaa` decomposes as `aaa`, then `baa`, then `ba`, then `baa`), which can be rewritten as

$$\mathcal{W} = \text{Seq}(\{a\}) \times \text{Seq}(b \times \text{Seq}(\{a\})).$$

Thus the generating functions equate as

$$W(z) = \frac{1}{1-z} \cdot \frac{1}{1-\frac{z}{1-z}} = \frac{1}{1-2z},$$

which is consistent with us having 2^n words of length n.

Example 163

Let $a^{<k}$ be shorthand for $\text{Seq}_{<k}(\{a\})$ (so one of the sequences of $0, 1, \cdots, k-1$ as); the generating function of this class is $1 + z + \cdots + z^{k-1} = \frac{1-z^k}{1-z}$. Thus, $\mathcal{W}^{(k)}$, the class of words without k consecutive appearances of the letter a, can be written

$$\mathcal{W}^{(k)} = a^{<k} \times \text{Seq}(b \times a^{<k}).$$

By the same logic, we thus have

$$W^{(k)}(z) = \frac{1-z^k}{1-z} \cdot \frac{1}{1 - \frac{z(1-z^k)}{1-z}} = \frac{1-z^k}{1-2z+z^{k+1}}.$$

To do asymptotics, we thus need the singularities of the denominator. And now we'll do yet a more involved example:

Example 164

Let $\mathcal{W}^{\alpha,\beta}$ be the number of words with at most α consecutive as and at most β consecutive bs.

This is a bit more involved than the previous examples: first, we decompose the set of all words into consecutive blocks of as and bs

$$\mathcal{W}^{\alpha,\beta} = \text{Seq}(b) \times \text{Seq}\Big(a \times \text{Seq}(a) \times b \times \text{Seq}(b)\Big) \times \text{Seq}(a).$$

To write $\mathcal{W}^{\alpha,\beta}$ in the same language as before, we take that expression above and replace $\text{Seq}(a)$ by $\text{Seq}_{<\alpha}(a)$ and $\text{Seq}(b)$ by $\text{Seq}_{<\beta}(b)$ inside the parentheses, and we also replace by $\text{Seq}_{\leq\alpha}(a)$ and $\text{Seq}_{\leq\beta}(b)$ outside the parentheses. Then we can plug in generating functions and simplify, and a computer can do the rest. For example if $\alpha = \beta = r$,

$$W^{r,r}(z) = \frac{1-z^{r+1}}{1-2z+z^{r+1}} = \frac{1+z+z^2+\cdots+z^r}{1-z-z^2-\cdots-z^r}.$$

The idea is that if you ask kids to write down a random binary sequence of length 200 and then write a computer program to generate such sequences, statisticians can tell which ones come from computers because humans intuitively expect much shorter consecutive sequences than random would produce! And analyzing this kind of generating function gets us things like that: the chance a random sequence of length n has a max consecutive sequence of length k is

$$\frac{1}{2^n}(W_n^{(k,k)} - W_n^{(k-1,k-1)}),$$

and this is bigger than 10 percent even when $n = 200, k = 11$.

Theorem 165

Any regular language has a rational generating function. (Indeed, it's a finite iteration of operations that we've seen which all preserve this property.) Therefore, the values of $L_n = |\mathcal{L}_n|$ will always satisfy some linear recurrence of the form

$$L_n = a_1 L_{n-1} + \cdots + a_k L_{n-k}, \quad a_i \in \mathbb{Q},$$

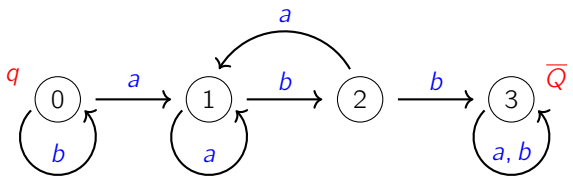
for all n sufficiently large and with some initial conditions.

This is only one of four different ways to describe all of this – we'll switch perspectives now to (finite) **automata**.

Definition 166

A **finite deterministic automaton** is a directed multigraph (loops and multiple edges allowed) with some finite vertex set Q and edge set E, where each edge is decorated with a symbol in some finite alphabet $\mathcal{A}$. We specify an initial state $q_0 \in Q$ and a set of terminal states $\overline{Q} \subseteq Q$, and here "deterministic" means that for every $q \in Q$ and every $\alpha \in \mathcal{A}$, there is at most one outgoing edge from q decorated with the symbol α.

The idea is that we start at the initial vertex, and then we're given a string in the alphabet which tells us what edges to follow. Here is an example:



Definition 167

A word $w = (w_1, \cdots, w_n)$ is **accepted** by an automaton A if there is a path starting at q_0 and ending at some state in $\overline{Q}$, such that the symbols along the path edges are $w_1, \cdots, w_n$ in that order.

We can check that a possible path in the automaton above is the word `ababba`, because we can just follow the symbols in the unique way possible: we are at states $0, 1, 2, 1, 2, 3, 3$.

Definition 168

A language is ***A*-regular** if there exists a deterministic finite automaton which accepts only the words in the language.

Example 169

The language accepted by our automaton drawn above is exactly the words containing the sequence `abb`.

Theorem 170 (Equivalence theorem)

A language is regular if and only if it is A-regular (and in fact there is a way of going back and forth between specifications of the two notions). There are other equivalences for random automata as well.

Theorem 171 (Chomsky-Schutzenberger)

Suppose G is a deterministic finite automata with $Q = \{q_0, \cdots, q_k\}$ where the initial state is q_0 and the final states are $\{q_{i_1}, \cdots, q_{i_a}\}$. Then the generating function for the language $\mathcal{L}$ of accepted words is a rational function given by the matrix expression

$$L(z) = U(I - zT)^{-1}V,$$

where $T_{ij} = 1\{\alpha \in A : \text{there is a path from } q_i \text{ to } q_j \text{ with symbol } \alpha\}$, $U = (1, 0, \cdots, 0)$, and $V = (v_0, \cdots, v_k)$ for $v_j = \delta_{\overline{Q}}(q_j)$ (that is, it's 1 for the terminal state).

There is a way of writing down this inverse in terms of the loops in the graph, and it's **Viennot's theory of heaps of pieces**. Don Knuth has a chapter in volume 4A dedicated to this if we're interested – the point is that if we can write down an automata for the combinatorial object, we can make some progress towards the generating function! (And in some classes, for example Dhar's directed animals, this is a natural way to approach the problem.)

Remark 172. *We did an example above with "words containing the pattern* `abb`*." We can do a similar thing for general patterns in a reasonable way, and here's a related story. Not all relations are transitive (for example "A beats B in a game"); for example, consider the magic square* $\begin{bmatrix} 4 & 3 & 8 \\ 9 & 5 & 1 \\ 2 & 7 & 6 \end{bmatrix}$*, and we have three piles of cards given by the three columns. If we each pick a pile and then randomly pick a card, the third column beats the second column with probability* $\frac{5}{9}$*, and the second column beats the first column with probability* $\frac{5}{9}$*, but the first column also beats the third column with probability* $\frac{5}{9}$*. And for the standard construction of* $n \times n$ *magic squares, the same "non-transitive game" situation occurs too. (There's a similar story we can tell with Efron's non-transitive dice.)*

Similarly, suppose we have a "penny game" where two people pick binary strings and see which one comes up first in a sequence of coin flips. To solve this problem and calculate winning probabilities, we need to be able to understand the rules for patterns, and the calculus of generating functions lets us do this carefully! It turns out 000 *and* 111 *do poorly against the other six strings, which split into two nontransitive triples. (For example, it's more likely for* 10 *to come up than* 00*, since as long as a* 1 *comes up at all,* 10 *will automatically win.) Writing down these generating functions uses Conway's algorithm, and there's lots of interesting ideas there too.*

25 May 28, 2025

Today, we'll (finally) describe how to take generating functions and use them to run the **Boltzmann sampler**. Throughout, we'll let $\mathcal{C}$ be a combinatorial class with weight function denoted $|\cdot|$: our goal will be to sample an object from $\mathcal{C}$ uniformly, meaning that we want $\gamma \in \mathcal{C}_n$ with $\mathbb{P}(\gamma) = \frac{1}{|\mathcal{C}_n|}$.

The idea (which is common in areas like statistical physics) is to take our generating function $C(x) = \sum_{n=0}^{\infty} C_n x^n$ and choose $\gamma \in \mathcal{C}$ from the infinite class with probability

$$\mathbb{P}_x(\gamma) = \frac{x^{|\gamma|}}{C(x)},$$

where we choose the parameter x so that $\mathbb{E}[|\gamma|] = n$. (In statmech this is called "going from the microcanonical to the canonical ensemble.") When we do this, we may not necessarily get the value of n that we want, but we can reject repeatedly until $\gamma \in \mathcal{C}_n$. Then indeed, given that $|\gamma| = n$, we'll be uniform on the class $\mathcal{C}_n$ and have the desired result. And as Flajolet once said, often if we want partitions of size 1000000, it can be useful to have partitions of size 1000010 or 999990 as well, so we don't necessarily need to discard everything else that we get.

Fact 173

Two basic papers in this topic are "Boltzmann samplers for the random generation of combinatorial structures" by Duchon, Flajolet, Louchard, and Schaeffer and "Boltzmann sampling of unlabelled structures" by Flajolet, Fusy, and Pivoteau; the Wikipedia page has some other good references as well. Everything we're doing here is with unlabeled structures, but there's also a parallel theory for labeled structures which uses the exponential generating function instead.

To make this lecture self-consistent, we'll mention some background on this class of distributions $\mathbb{P}_x(\gamma) = \frac{x^{|\gamma|}}{C(x)}$ – these are called "exponential families" or "Boltzmann distributions," and of course this generating function must converge for this to make sense (so we need some bounds on the radius of convergence).

Lemma 174

Let N be the weight of the object sampled under this measure. We have

$$\mathbb{E}_x[N] = \frac{xC'(x)}{C(x)}, \quad \mathrm{Var}_x(N) = \frac{x^2C''(x) + xC(x)}{C(x)} - \left(\frac{xC'(x)}{C(x)}\right)^2.$$

Proof. The generating function of N is

$$\sum_{n=0}^{\infty} P_x(N = n)z^n = \frac{C(xz)}{C(x)},$$

so if we differentiate both sides in z we get

$$\mathbb{E}_x[N] = \frac{\partial}{\partial z}\left(\frac{C(xz)}{C(x)}\right)_{z=1} = \frac{xC'(x)}{C(x)}$$

by the chain rule as desired. Similarly we have

$$\mathbb{E}_x[N(N-1)] = \frac{\partial^2}{\partial z^2}\left(\frac{C(xz)}{C(x)}\right)_{z=1},$$

and then doing out the derivatives and using that $\mathrm{Var}(N) = \mathbb{E}[N^2] - \mathbb{E}[N]^2$ yields the result. □

As a consequence, $\mathbb{E}_x[N]$ is increasing in x (since $x\frac{d}{dx}\mathbb{E}_x[N] = \mathrm{Var}_x(N)$ is positive), and thus we can sample larger values by increasing x toward the radius of convergence.

Everything we talk about here generalizes to multiple parameters as well: we can have vector-valued weights and the theory still works.

Example 175

Consider binary words, where $|C_n| = 2^n$ and thus $C(x) = \frac{1}{1-2x}$. The radius of convergence of this power series is $x = \frac{1}{2}$; when $x = 0.4$ we have $\mathbb{E}_x[N] = 4$ and $\mathrm{SD}(N) \approx 4.47$, and when $x = 0.49505$ we have $\mathbb{E}_x[N] = 100$ and $\mathrm{SD}(N) \approx 100.5$. So when we do rejection sampling, we basically need to do 100 different samples before we get a binary word of the correct length. But in some other examples we get something better.

Recall that we build up regular languages with some basic operations, and it turns out we can associate those to some basic samplers. Throughout this next discussion, for $\mathcal{A}$ a class we'll let $\Gamma A(x)$ be the random variable outputted by the sampler.

- If $\mathcal{A} = \{f_1, \cdots, f_n\}$ is a class with finitely many elements, then we select b_i with probability

$$\frac{x^{|b_i|}}{\sum_j x^{|b_j|}}.$$

  So as long as we don't have too many elements, this is tractable.

- (Disjoint union) Suppose $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, and $C(x) = A(x) + B(x)$. So we have a mixture of two different classes, and for any fixed x let $p = \frac{A(x)}{A(x)+B(x)}$. We then flip a p-coin; if it comes up heads then we output $\Gamma A(x)$ and

otherwise we output $\Gamma B(x)$. Write the output of this as

$$\left(\text{Bin}\left(\frac{A(x)}{C(x)}\right) \to \Gamma A(x) \middle| \Gamma B(x)\right).$$

Here "bin" stands for binary, and we can do the same thing with a finite set of options to choose between; we write that as

$$(p_1, \cdots, p_n \to \Gamma A_1(x) | \cdots | \Gamma A_n(x)).$$

- (Products) For products of classes we have $C(x) = A(x)B(x)$, and $\Gamma C(x)$ can just return the ordered pair $(\Gamma A(x), \Gamma B(x))$, which we also write as $\Gamma A(x); \Gamma B(x)$.
- (Sequences) Suppose $\mathcal{C} = \text{Seq}(\mathcal{A})$, so that $C(x) = \frac{1}{1-A(x)}$. Recall that the size of the sequence is the sum of the sizes of the components. We can sample from this using the following geometric strategy: draw k from the $\text{Geom}(A(x))$ distribution, where the probability that a $\text{Geom}(\lambda)$ random variable is equal to n is $(1-\lambda)\lambda^n$. Then independently sample k times according to the procedure $\Gamma A(x)$ and return the resulting k-tuple. Notice that in order to do this, we must have $A(x) < 1$.

  We can also phrase this in a recursive way and define $\Gamma C(x)$ as follows: we flip a coin with probability $A(x)$; if it comes up heads then we return $(\Gamma A(x), \Gamma C(x))$ and otherwise we stop.

These are the main three samplers of the day, and we can do quite a lot of random generation with this (remember the languages we can build up using sums, products, and sequences are exactly the regular languages with rational generating functions).

Example 176

Professor Diaconis's handout has some figures from the two papers mentioned before: we can sample constrained binary words, random unbalanced $2-3$ trees, random connected non-crossing graphs, and plane partitions of fairly large sizes. For example,

$$\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}, \quad \mathcal{R} = \text{Seq}^{m,m}(\mathcal{A})$$

is the set of all words on the alphabet of two letters with no run of length greater than m.

We described how to get the generating function earlier in Example 164 – we do so by modifying the class of all words so that each consecutive run has at most size m, and this gives us a generating function. So because this is one of those expressions that we can build up, we can also build up the sampler, and here's how it's done: $\Gamma R(x)$ can be written as

$$\{X \implies \mathtt{b}\}; \Gamma\text{Cor}(x); \{X' \implies \mathtt{a}\},$$

where

$$\Gamma\text{Cor}(x) = \text{Geom}\left(\frac{x^2(1-x^m)^2}{(1-x)^2}\right) \implies (Y \implies \mathtt{a}; Y' \implies \mathtt{b})$$

for X, X' independent $\text{Geom}_{\{0,1,\cdots,m\}}(x)$ random variables and Y, Y' independent $\text{Geom}_{\{1,2,\cdots,m\}}(x')$, and where $Y \implies \mathcal{D}$ means that we pick an integer m from Y and then pick m iid copies of $\mathcal{D}$. Here $\frac{x(1-x^m)}{1-x}$ is the generating function for a geometric on $\{1, \cdots, m\}$, and we square that value because we have to run two of them in the "Cor" step.

Example 177

For the case $m = 4$ (so we want to generate things with no five consecutive symbols the same) and $x = 0.5$, the expected size of the random word that we get is 27. When this sampler was run three times, the resulting sequence had lengths $123, 23, 35$. If we wanted sequences of length 100, we'd increase x, but then we'd still need to do a lot of iterations. And there's lots of discussion about how exactly to choose the mean to maximize the acceptance probability.

Remark 178. *However, it takes several hours to try to generate a partition in this way if we tune x so that the mean length is 1 million. And so this is why the Burnside process comes in and saves the day – sometimes this sampler works well, and sometimes we don't need to care so much. We'll see some of the other ideas in the coming lecture.*

Definition 179

Suppose we have a system of equations $\mathcal{L}_1 = \phi_1(\mathcal{L}_1, \cdots, \mathcal{L}_n)$, $\mathcal{L}_2 = \phi_2(\mathcal{L}_1, \cdots, \mathcal{L}_n)$, and so on, up to $\mathcal{L}_n = \phi_n(\mathcal{L}_1, \cdots, \mathcal{L}_n)$, where all ϕ_is are polynomials (meaning they involve only unions and products). We then call $\mathcal{L}_1$ a **context-free language**.

The point is that in any such construction, there is some polynomial $f \in \mathbb{Q}[x, y]$ so that $L_1(x)$ satisfies $f(x, L_1(x)) = 0$, and therefore we have a generating function of algebraic type.

Example 180

Let $\mathcal{B}$ be the family of rooted binary trees. Then we know that $\mathcal{B} = \mathcal{Z} + (\mathcal{Z} \times \mathcal{B} \times \mathcal{B})$, which we can solve to find $B(x) = x + xB(x)^2 \implies B(x) = \frac{1-\sqrt{1-4x^2}}{2x}$. So the process $\Gamma B(x)$ which generates a random binary tree is performed as follows: flip a coin with probability $\frac{x}{B(x)}$; if it is heads, then we halt with just the root, and otherwise we call the process recursively for each of the left and right children.

In the language of the paper, we thus write this as

$$\Gamma B(x) = \left(\text{Bin}\left(\frac{x}{B(x)}\right) \implies \mathcal{Z} \middle| \mathcal{Z}; \Gamma B(x); \Gamma B(x); \right).$$

And this is basically a Galton-Watson branching process: at each vertex we have either 0 or 2 children with probabilities $\frac{x}{B(x)}$ and $1 - \frac{x}{B(x)}$, and this will eventually die out if $\frac{x}{B(x)} > \frac{1}{2}$.

26 May 30, 2025

We'll add a tool to our toolbox today and deal with **multisets** for the Boltzmann sampler. Suppose as usual that $\mathcal{A}$ is a combinatorial class with size function $|\cdot|$, and we let $\mathcal{A}_n$ be the subset of $\mathcal{A}$ of elements of size n. We can then take the class of finite multisets $\mathcal{M} = \text{MSet}(\mathcal{A})$, which we can write as

$$(\alpha_1, n_1), (\alpha_2, n_2), \cdots, (\alpha_\ell, n_\ell)$$

for $\alpha_i \in \mathcal{A}$ and n_i the multiplicity of each α_i appearing. Recall that the generating function can then be computed by noting that

$$\text{MSet}(\mathcal{A}) = \prod_{\alpha \in \mathcal{A}} \text{Seq}(\alpha) \implies M(z) = \prod_{\alpha \in \mathcal{A}} \frac{1}{1 - z^{|\alpha|}} = \prod_{n=1}^{\infty} \frac{1}{(1 - z^n)^{A_n}},$$

and so we can exponentiate and write this as $\exp\left(\sum_{j=1}^{\infty} \frac{A(z^j)}{j}\right)$. The object is to figure out how to generate from this generating function.

Example 181

Consider the set $\mathcal{P}$ of all integer partitions. A partition of size n can be written as $1^{n_1}2^{n_2}\cdots k^{n_k}$ with $\sum_i i n_i = n$, and (recalling that $\mathcal{I}$ is the class of positive integers)

$$\mathcal{P} = \text{MSet}(\mathcal{I}).$$

Since $I(z) = \frac{z}{1-z}$ (because each integer j has weight j), we can plug this in and find the generating function for partitions

$$P(z) = \prod_{n=1}^{\infty} \frac{1}{1-z^n} = \exp\left(\sum_{j=1}^{\infty} \frac{z^j}{j(1-z^j)}\right).$$

Similarly for any subset $\Omega \subseteq \mathcal{I}$, we can let $\mathcal{P}_\Omega$ be the set of all partitions with parts only in Ω. We then replace $\frac{z}{1-z}$ with whatever the generating function for Ω is, and that will yield some other result.

Example 182

Let $\mathcal{T}$ be the set of all unlabeled rooted (Polya) trees. Each tree can be broken up into the root and its parts, so

$$\mathcal{T} = \mathcal{Z} \times \text{MSet}(\mathcal{T})$$

yields a recursive formula

$$T(z) = z\exp\left(\sum_{j=1}^{\infty} \frac{T(z^j)}{j}\right).$$

We can also define $\mathcal{T}_\Omega$ to be the set of trees whose outdegrees are in Ω, and similarly we can easily get the generating function. So in summary, we've done various constructions of combinatorial objects, and the question of the day will be how to sample from $\text{MSet}(\mathcal{A})(x)$ (that is, sample a multiset γ with probability $\frac{x^{|\gamma|}}{\text{MSet}(\mathcal{A}(x))}$) given a sampling method for $\mathcal{A}$. We'll first write down the algorithm and then explain it:

- Define a probability measure on $\{1, 2, 3, \cdots\}$ via the cumulative distribution function

$$\mathbb{P}(K = k) = \frac{\prod_{j\le k} \exp\left(\frac{1}{j}A(x^j)\right)}{Z}, \quad Z = \prod_{j=1}^{\infty} \exp\left(\frac{1}{j}A(z^j)\right).$$

  (This approaches 1 as $k \to \infty$, so it's a valid probability distribution.) Let k_0 be a sample from this distribution.

- Now initialize γ to the emptyset and perform the following for each j from 1 to $(k_0 - 1)$:

$$\gamma \leftarrow \gamma, \left[\text{Pois}\left(\frac{A(x^j)}{j}\right) \implies \text{Copy}(j, \Gamma A x^j)\right].$$

  That is, sample from A with parameter x^j, put j copies of that object in our multiset γ, and do that process a Poisson number of times.

- Finally, do the same for $j = k_0$ but condition the Poisson random variable to be positive – that is,

$$\gamma \leftarrow \gamma, \left[\text{Pois}_{\ge 1}\left(\frac{A(x^{k_0})}{k_0}\right) \implies \text{Copy}(k_0, \Gamma A x^{k_0})\right].$$

- Finally, γ is the multiset we return.

This gives us a random variable, and it turns out this variable can be used to prove some interesting theorems! But what we'll do now is understand why it works.

Proof of correctness. By definition we have

$$\mathcal{C} = \text{MSet}(\mathcal{A}) = \prod_{\alpha \in \mathcal{A}} \text{Seq}(\alpha),$$

and so $C(z) = \prod_{\alpha \in \mathcal{A}} \frac{1}{1-z^{|\alpha|}} = \prod_{j=1}^{\infty} \exp\left(\frac{A(z^j)}{j}\right)$. So heuristically the idea is that picking from the multiset is the same as sampling each α with multiplicity $\text{Geom}(x^{|\alpha|})$, and that would be a valid algorithm if our class were finite.

Instead what we do is **convert the geometric to a Poisson**:

Lemma 183

Let $\{Y_i\}_{i=1}^{\infty}$ be independent Poissons of parameter $\frac{\lambda^i}{i}$ (these will eventually be zero by Borel-Cantelli). If $N = \sum_{i=1}^{\infty} iY_i$, then $N \sim \text{Geom}(\lambda)$.

Proof of lemma. It's easy to verify that this is true by comparing the Laplace transforms or generating functions of both variables, but recall that we've already proved this: we have the formula for the cycle index

$$Z_{S_n}(x_1, \cdots, x_n) = \sum_{\sigma \in S_n} \prod_{i=1}^{N} x_i^{a_i(\sigma)} = \sum_{\lambda \vdash n} \frac{1}{Z_\lambda} \prod_i x_i^{a_i(\lambda)}, \quad Z_\lambda = \prod_i i^{a_i} a_i!,$$

and we proved very early in the quarter that

$$\sum_n t^n Z_{S_n}(x_1, \cdots, x_n) = \exp\left(\sum_{i=1}^{\infty} x_i \frac{t^i}{i}\right).$$

Multiplying both sides by $(1-t)$ then yielded

$$\sum_n t^n (1-t) Z_{S_n}(x_1, \cdots, x_n) = \exp\left(\sum_{i=1}^{\infty} (x_i - 1)\frac{t^i}{i}\right),$$

where the right-hand side is the generating function of a Poisson with parameter $\frac{t^i}{i}$ and the left-hand side is a geometric with parameter t. So what this means is that picking n from $\text{Geom}(1-t)$ and then picking $\sigma \in S_n$ uniformly has cycle types independent Poissons with parameter $\frac{t^i}{i}$ for length-i. □

So turning back to the multiset $\Gamma C(x)$ we want to build, we can instead think of

$$\begin{aligned}
\Gamma C(x) &= \prod_{\alpha \in \mathcal{A}} \alpha^{\sum_{i=1}^{\infty} i\text{Pois}\left(\frac{x^{|\alpha|i}}{i}\right)} \\
&= \prod_i \prod_{\alpha \in \mathcal{A}} \alpha^{i\text{Pois}\left(\frac{x^{|\alpha|i}}{i}\right)} \\
&= \prod_i \prod_{\beta \in \mathcal{A}^{\otimes i}} \beta^{\text{Pois}\left(\frac{x^{|\beta|i}}{i}\right)},
\end{aligned}$$

where $\mathcal{A}^{\otimes i}$ is like $\mathcal{A}$ but each element is replaced with itself repeated i times. So now give any class $\mathcal{B}$ with some size function, we can let $\mathcal{P}$ be the sampling problem $\prod_{\beta \in \mathcal{B}} \beta^{\text{Pois}(cx^\beta)}$ for some $c > 0$ and $x < 1$. If we want to pick these

independent Poissons for each β, this can be realized by the easier sampling problem

$$\mathcal{U} : [\text{Pois}(cB(x)) \implies \Gamma B(x)].$$

That is, pick a Poisson with parameter $cB(x)$ and then iid sample from B that many times. The reason this works is that the probability $\mathcal{P}$ results in a multiset $\gamma_1^{r_1}\gamma_2^{r_2}\cdots\gamma_s^{r_s}$ is (plugging in the probability mass function for the Poisson and using our lemma)

$$\prod_{i=1}^{s} \frac{\left(cx^{|\gamma_i|}\right)^{r_i}}{r_i!} \prod_{\beta\in\mathcal{B}} e^{cx^{|\beta|}} = c^{\ell}x^{N}e^{cB(x)} \prod_{i=1}^{s} \frac{1}{r_i!},$$

where $\ell = \sum r_i$ and $N = \sum |\gamma_i| r_i$, and then if we try sampling from $\mathcal{U}$ instead we also get this exact same expression because we can end up with the same multiset in any of $\binom{\ell}{r_1,\cdots,r_s}$ possible ways, and the rest of the factors exactly work out.

So if we apply this to our multiset problem $\Gamma C(x)$, the easier sampling problem tells us that for each "repetition size" i we can sample from $\prod_{\beta\in\mathcal{A}^{\otimes i}} \beta^{\text{Pois}\left(\frac{x^{|\beta|i}}{i}\right)}$ with this Poissonized strategy instead. This is still an infinite product, but we can also calculate ahead of time the distribution of k_0, the largest size j which will have a nonzero Poisson $\frac{A(x^j)}{j}$. Indeed, $\mathbb{P}(K \le j)$ is the probability that all Poissons past j are equal to zero, which is the product $\exp\left(-\sum_{i\ge j} A\left(\frac{x^i}{i}\right)\right)$. That's exactly how we chose our probability distribution for k_0 above. □

Remark 184. *We can indeed sample from this distribution K because we can calculate $\mathbb{P}(K = k) = \mathbb{P}(K \le k) - \mathbb{P}(K \le k-1)$; even if we can't calculate Z exactly we can truncate after some large number of terms. And so we also need the generating function $A(x^j)$ and to be able to sample from As (or else things are numerically challenging), but other than that everything here is completely routine!*

Fact 185

One small question is to ask what this algorithm becomes for partitions $\mathcal{P}$; there's a standard algorithm which uses geometric random variables, and it's worth double-checking whether this algorithm reduces to the exact same thing. When we try generating partitions of size 1000 with code, it unfortunately does take a lot of repetitions (and many hours) to get exactly 1000.

For another example, suppose we want to generate Polya trees. Our story is that $\mathcal{T} = \mathcal{Z} \times \text{MSet}(\mathcal{T})$, and so if we do this algorithm it becomes a self-referential one! There's some steps to fill in, but this does turn out to still be useful if we're slightly more careful ("either we end or we pick the degree below from some computable distribution"). For details, we can see "Scaling limits of random Polya trees" by Panagiotou and Stufler.

As a final comment, if we want to generate powersets (that is, random subsets with no repeats) of $\mathcal{A}$ with probability proportional to $x^{|S|}$, there is a trick:

$$\text{MSet}(\mathcal{A}) = \text{PSet}(\mathcal{A}) \times \text{MSet}(\mathcal{A}^{\otimes 2}),$$

since all repetitions are either even or odd in length, and so what we can do is just sample multisets and then keep an element if and only if it shows up an odd number of times. (And for more details, we see the paper by Flajolet, Fusy, and Pivoteau.)

27 June 2, 2025

Fact 186

I was out of town for the last two lectures of the course, so these notes are transcribed from class lecture notes. Apologies in advance for any errors I've introduced during this process!

The last topic of the quarter will be **conditioned limit theory and equivalence of ensembles**, and the idea (as previously mentioned) is to go from our algorithms for building generating functions to proving theorems. We know that for the combinatorial classes $\mathcal{C}$ we've been iteratively building, we have ways of also building the corresponding generating functions $C(x) = \sum_n C_n x^n$. But often we also want to understand $\mathcal{C}_n$, the set of elements of size n, or $C_n = |\mathcal{C}_n|$, the number of such elements.

Example 187

We'll illustrate the problem here by considering **Abel summability**. The setting is that we have a sequence $\{a_n\}_{n=0}^{\infty}$ of real numbers, and we can consider the ordinary limit $a_n \to \ell$, the **Cesaro convergence** $\frac{1}{n}\sum_{j=0}^{n-1} a_j \to \ell$, and the **Abel convergence** $(1-x)\sum_{n=0}^{\infty} a_n x^n \to \ell$.

These often agree, and in fact we have some easy implications:

Proposition 188 ("Abelian theorem")

If $a_n \to \ell$, then $a_n \overset{\text{Ces}}{\to} \ell$ as well, and if $a_n \overset{\text{Ces}}{\to} \ell$, then $a_n \overset{\text{Abel}}{\to} \ell$.

On the other hand, the converses are not true. Indeed, the sequence $a_n = (-1)^n$ does not converge but does converge in the Cesaro sense (since $\frac{1}{n}\sum_{j=0}^{n-1}(-1)^j \to 0$), and similarly it converges in the Abel sense because $(1-x)\sum_{n=0}^{\infty}(1-x)^n = \frac{1-x}{1+x} \to 0$. Similarly, $a_n = (-1)^j j$ does not have a limit and also does not converge in the Cesaro sense (because $-1+2-3+4-\cdots+2n = n$ but $-1+2-3+\cdots+(2n-1) = -n$), but it does converge in the Abel sense via the calculation

$$\sum_{j=0}^{\infty}(-1)^j j x^j = -\frac{x}{(1+x)^2} \implies (1-x)\sum a_j x^j = -\frac{x(1-x)}{(1+x)^2} \to 0.$$

So it's an interesting question to ask **what conditions guarantee the converses**; that is, conditions on a_n so that Abel convergence implies Cesaro or ordinary convergence. These are called **Tauberian conditions** and are a healthy subject (for details, we can see Hardy's book "Divergent series," Feller volume II, or Professor Diaconis' paper "G.H. Hardy and Probability???"); for example if the sequence is nonnegative then Cesaro convergence implies ordinary convergence, and for monotone sequences this is equivalent to ordinary convergence.

Convergence of the series $C(x) = \sum C_n x^n$ can be more complicated than this, though – of course there might be more complicated singularities. A main concept of the book by Flajolet and Sedgewick is to use the behavior of $C(x)$ to get information about its coefficients via complex analysis, and in these lectures we'll see how we can actually get this via probability instead!

Example 189

Consider the case of $\mathcal{P}$, the class of integer partitions. Recall that we have $\mathcal{P} = \text{MSet}(\mathcal{I})$, which yields $P(z) = \prod_{j=1}^{\infty}(1-z^j)^{-1}$ and therefore $\mathcal{P} = \prod_j j^{\text{Geom}(x^j)}$.

Last lecture, we "massaged this" to get what we wanted, but we don't have to: we can just pick A_j independently

from Geom(z^j) and let $\lambda = 1^{a_1}2^{a_2}\cdots$, and that makes sense on its own. To match classical notation, we'll write $z = q$ from here on for some $0 < q < 1$.

Proposition 190

Consider the probability distribution Q_q on $\mathcal{P}$ given by

$$Q_q(\lambda) = q^{|\lambda|}\prod_{j=1}^{\infty}(1-q^j).$$

(Indeed, if we sum over all λ of a given size we get $p(n)q^n\prod_{j=1}^{\infty}(1-q^j)$, and then summing over n yields 1.) Then under Q_q, if the partition sampled is $\lambda = 1^{A_1}2^{A_2}\cdots$, then A_k are Geom(q^k) (meaning that $\mathbb{P}(A_k = j) = q^{kj}(1-q^k)$.

In particular, this means that if $N = \sum_j jA_j$ for random variables of this type, then $\mathbb{P}(N = n) = p(n)q^n\prod_{j=1}^{\infty}(1-q^j)$, and then for any subset of partitions $A \subseteq \mathcal{P}_n$, we have $P_n(A) = Q_q(A|N = n)$. If we differentiate like we did last lecture "on N" (that is, on the sum $\sum jA_j$), then

$$\mathbb{E}_q[N] = \sum\frac{kq^k}{1-q^k}, \quad \mathrm{Var}(N) = \sum\frac{k^2q^k}{(1-q^k)^2}.$$

Lemma 191

We have the asymptotic expression $\mathbb{E}_q[N] = n + O(n^{3/4})$ if we set $q = q_n = e^{-\pi/\sqrt{6n}}$.

The proof sketch of this is to take our expression $\sum\frac{kq^k}{1-q^k}$ and rewrite it as a Riemann sum

$$\frac{1}{\log^2(1/q)}\int_0^{\infty}\frac{ue^{-u}}{1-e^{-u}} + O\left(\frac{1}{\log 1/q}\right) = \frac{\pi^2}{6\log^2(1/q)} + O\left(\log\frac{1}{q}\right).$$

We can then choose $q_n = e^{-\pi/\sqrt{6n}}$ to get $Q_{q_n}(N = n) \sim \frac{1}{\sqrt[4]{96n^3}}$, which means that it takes on the order of $n^{3/4}$ trials to get a partition of the size that we want. (For something like $n = 10^6$, this requires roughly 100000 samples, and we can compare the efficiency of that with something like the Burnside process which is empirically much faster!)

But we can get much more – Fristedt's work (from 1993) uses conditioned limit theory to get results such as

$$\mathbb{P}\left(\frac{\pi}{\sqrt{6n}}kA_k \le r\right) \sim 1 - e^{-r}$$

(meaning that $A_k \sim \frac{c\sqrt{n}}{k}$) and that the A_ks up to $O(\sqrt{n})$ are independent. We can also get that the large parts satisfy $P_n\left(\frac{\pi}{\sqrt{6n}}Y_1 - \log\frac{\sqrt{6n}}{\pi} \le r\right) \to e^{-e^{-r}}$, that the number of parts is $c\sqrt{n}\log n$, and so on. So let's do an intro to this and start to understand all of the theory.

The setting is as follows: we let N_i be the "number of things of type i" for $1 \le i \le k$, and we sometimes have the situation that

$$\mathbb{P}(N_1 = n_1, \cdots, N_k = n_k) = \mathbb{P}(X_1 = n_1, \cdots, X_k = n_k | T(x_1, \cdots, x_k) = N)$$

for **independent** random variables X_i. In our example, we had X_j = Geom(q^j) and the function $T(x_1, \cdots, x_k) = \sum_j jX_j$. And here's another worked out example with more details:

Example 192

One of Professor Diaconis' friends' kid came home one day with a lot of money. The setting is as follows: consider a shuffled deck of 52 cards and cut off the top 26 cards. We say that a "book" occurs if a half has 4 of a kind – the kid offered 3-to-1 odds that there is no book in either half.

We wish to compute what the fair odds are, and here's how we'll do so. Let N_i be the number of cards of value i in the top 26 cards, so that $\mathbb{P}(\text{no book}) = \mathbb{P}(1 \le N_i \le 3 \text{ for all } 1 \le i \le 13)$. The key observation is that for any $0 < p < 1$, we can let X_i be iid $\text{Bin}(4, p)$ random variables, and then

$$\mathbb{P}(\text{no book}) = \mathbb{P}\left(1 \le X_i \le 3 \text{ for all } i \middle| \sum_{i=1}^{13} X_i = 26\right).$$

This expression can in fact be computed using a Bayes' theorem trick: in general we can write

$$\begin{aligned}\mathbb{P}(a_i \le N_i \le b_i \text{ for all } i) &= \mathbb{P}(a_i \le X_i \le b_i | T(\vec{x}) = t) \\ &= \mathbb{P}(T(\vec{x}) = t | a_i \le X_i \le b_i \text{ for all } i) \cdot \frac{\mathbb{P}(a_i \le X_i \le b_i \text{ for all i})}{\mathbb{P}(T(\vec{x}) = t)},\end{aligned}$$

and now all three terms in this last expression only involve independent random variables and thus can be easily computed! In our setting, if we take X_i to be $\text{Bin}(4, \frac{1}{2})$, then conditioned on $1 \le X_i \le 3$ the random variable takes on values $1, 2, 3$ with probability $\frac{2}{7}, \frac{3}{7}, \frac{2}{7}$. So the probability of no book in our case is

$$\mathbb{P}\left(\sum_{i=1}^{13} X_i = 26 \middle| 1 \le X_i \le 3 \text{ for all } i\right) \cdot \frac{\mathbb{P}(1 \le X_i \le 3)^{13}}{\mathbb{P}(\sum_{i=1}^{13} X_i = 26)},$$

which we can approximate as $\left(\frac{7}{8}\right)^{13} \frac{\sigma_X}{\sigma_Y}$ for Y the conditioned version of X; this yields 0.2331 (and the correct answer is 0.23145), which is much higher than $\mathbb{P}(1 \le X_i \le 3 \text{ for all } i) = 0.1762$. So the "fair odds" are 3.311, and the kid was earning more than 4 cents per dollar!

Example 193

For another example calculation involving balls in urns, suppose an urn has B_i balls of color i and $\sum_i B_i = B$. If we sample $n < B$ balls without replacement, then

$$\mathbb{P}(N_i = a_i \text{ for all } i) = \mathbb{P}\left(X_i = a_i \text{ for all } i \middle| \sum_i X_i = n\right)$$

if we sample X_i to be $\text{Bin}(B_i, p)$ for any fixed constant p.

The point is that such **conditional representations** show up quite a lot – there are dozens of them and we'll see a list of some next lecture. The symbolic machine that we've gone through actually has a hundred more of them (though not much work has been done to use them)! And somehow the idea is that "limit theorems and simulations are complementary" – being able to run simulations also yields some limiting results and vice versa.

28 June 4, 2025

We'll start by listing off some of these various conditioned limit results that we alluded to last time:

1. Urns, multinomials, and the Bayes trick (as we did in the example last time): Bruce Levin's "A Representation for Multinomial Cumulative Distribution Functions."
2. Partitions of n and geometric random variables (also mentioned last time): Bert Fristedt's "The Structure of Random Partitions of Large Integers."
3. Compositions of n: Diaconis, Holmes, Janson, Lalley, and Pemantle's "Metrics on Compositions and Coincidences among Renewal Sequences."
4. Prime numbers and the zeta function: Professor Diaconis' "Average running time of the fast Fourier transform."
5. Points on high-dimensional spheres and spacings: Diaconis and Freedman's "A dozen de Finetti-style results in search of a theory," as well as Ronald Pyke's "Spacings."
6. Set partitions (Bell numbers): Bert Fristedt's "The structure of random partitions of large sets," as well as Chern, Diaconis, Kane, and Rhoades' "Central Limit Theorems for some Set Partition Statistics."
7. Factoring polynomials and riffle shuffling: Diaconis, McGrath and Pitman's "Riffle shuffles, cycles, and descents."
8. Le Cam's method and applications: Lars Holst's "Two Conditional Limit Theorems with Applications."
9. Random matrix theory: Jason Fulman's "Random matrix theory over finite fields: a survey."
10. Random permutations: Shepp and Lloyd's "Ordered Cycle Lengths in a Random Permutation."
11. Finally, connections with equivalence of ensembles and exchangeability (de Finetti's theorem): Diaconis and Freedman's "Partial exchangeability and sufficiency."

The point of this last lecture is to do various examples (and the literature review mentioned above) of conditional limit theorems.

Example 194

Recall that we've already done an elaborate example earlier in the course with random permutations and wreath products: we have the cycle index formula

$$Z_n(x_1, \cdots, x_n) = \frac{1}{n!} \sum_{\sigma \in S_n} \prod_i x_i^{a_i(\sigma)} \implies \sum_{n=0}^{\infty} t^n (1-t) Z_n = \exp\left(\sum_i (x_i - 1) \frac{t^i}{i} \right).$$

This says that if we pick N from a geometric distribution and then $\sigma \in S_N$ uniformly, then writing σ in cycle notation yields independent Poissonian a_is with parameters $\frac{t^i}{i}$.

Example 195 (Factoring polynomials over $\mathbb{F}_q$)

Let $q = p^\Delta$ be a prime power, and pick one of the q^n monic degree-n polynomials in $\mathbb{F}_q[x]$. We can then factor f into monic irreducible polynomials and let N_i be the number of factors of degree i.

It turns out we have the following conditional limit theorem in this case:

Theorem 196

We have

$$P_n(N_1 = n_1, \cdots, N_k = n_k) = \mathbb{P}\left(X_1 = n_1, \cdots, X_k = n_k \middle| \sum_{i=1}^{n} i x_i = n\right),$$

where X_i are independent and have the distributions

$$X_i \sim \text{NegBin}\left(f_{iq}, q^{-i}\right), \quad f_{iq} = \frac{1}{i}\sum_{d|i} \mu(d) q^{i/d}.$$

In particular, this means that questions about the number of factors, the largest factor, the number of linear factors, and so on, are all in control for factorization of a uniform random polynomial.

This example turns out to be related to riffle shuffles: if we shuffle n cards ℓ times via riffle shuffles from the Gilbert-Shannon-Reeds model, that yields a measure on the permutation group S_n, which we can factor into cycles.

Theorem 197

After ℓ Gilbert-Shannon-Reeds shuffles on n cards, the cycle distribution of the resulting permutation $\mathbb{P}(a_1(\sigma) = m_1, \cdots, a_n(\sigma) = m_n)$ has exactly the same distribution as in the theorem above.

There is more known about why these expressions are equal – we can see Professor Diaconis' survey "Mathematical developments from the analysis of riffle shuffling" for more.

Example 198

Next, pick a uniform matrix from $\text{Mat}_{n\times n}(q)$ (or $\text{GL}_n(q)$, but we'll stick to all matrices for now). We are interested in understanding the number of fixed vectors, the numbers and sizes of Jordan blocks, and so on.

In this case, the object to study is the characteristic polynomial – as usual, we can factor it into monic irreducibles of various degrees and multiplicities. From this, we can use **rational canonical form**: take each irreducible polynomial $\phi = \sum a_j x^j$ and construct its companion matrix

$$C(\phi) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{m-2} & -a_{m-1} \end{bmatrix},$$

which can then be put together into diagonal blocks. The data needed to specify this form are $\{p_i, \lambda_i\}$ (an integer partition associated to each irreducible), where $\sum |\lambda_i| \deg(p_i) = n$, and our question is "what p_i and λ_i do we choose?". As before, the story is to randomize n as $\mathbb{P}(N = n) = (1-u)u^n$ and then pick $M \in \text{Mat}_{n\times n}$ uniformly over all choices. It turns out once we do this, the partitions $\{\lambda_p\}$ are independent with "nice law" involving q and the corresponding polynomial p.

Remark 199. *There are many more examples where this kind of randomization of a parameter makes the components independent, but of course these are all special in some way (we can think of them as being the "exactly solvable models").*

Example 200

Turning now to prime factors, suppose we pick an integer n uniformly at random from $\{1, 2, \cdots, N\}$ and want to understand how many primes divide n. More generally, instead of considering $\sum_{p|n} 1$, we might be interested in $\sum_{p|n} p$, or the largest prime factor, or some other related quantity.

Part of the intuition here is that "half of all numbers are even and a sixth of the numbers are divisible by 6," so $\mathbb{P}(2 \text{ and } 3|n) = \mathbb{P}(2|n)\mathbb{P}(3|n)$ and so on (at least approximately); this logic also yields that $\mathbb{P}(n \text{ squarefree}) = \prod_p \left(1 - \frac{1}{p^2}\right) = \frac{6}{\pi^2}$.

To make this more precise, we'll define a probability measure on the positive integers

$$\mathbb{P}(j) = \frac{1}{\zeta(2)} \frac{1}{n^2}, \quad \zeta(2) = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Under this measure, when we sample an integer N, we have $\mathbb{P}(m|N) = \frac{1}{\zeta(2)} \sum_{k=1}^{\infty} \frac{1}{(mk)^2} = \frac{1}{m^2}$, so that if $N = \prod p^{a_p(N)}$, the exponents $a_p(N)$ are independent, with $\mathbb{P}(a_p(N) = 0) = 1 - \frac{1}{p^2}$, and more generally $\mathbb{P}(a_p(N) = j) = \left(1 - \frac{1}{p^2}\right) \frac{1}{p^{2j}}$. So for quantities like the ones mentioned above, we can calculate

$$\omega(n) = \sum_p 1\{a_p \geq 1\}$$

and similarly we can study quantities like $\sum_p a_p$. For the number of prime factors, we do have a well-known result:

Theorem 201 (Hardy-Ramanujan-Erdos-Kac)

Letting $\omega(n)$ be the number of distinct prime factors of n,

$$\mathbb{P}_N \left(\frac{\omega(n) - \log\log n}{\sqrt{\log\log n}} \leq x \right) \to \Phi(x).$$

Example 202

We'll next turn to the general **Le Cam's method**. The idea is as follows: we have independent random variables X_i (plus some additional conditions) and wish to study a conditional law $\mathcal{L}(\sum_{i=1}^n g_n(X_i) | \sum_{i=1}^n X_i = t)$.

The rough idea is that if we can show

$$\begin{bmatrix} a_n(\sum g_n(x_i) - \mu_n) \\ b_n(\sum X_i - \nu_n) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} Y \\ X \end{bmatrix},$$

then we have that $\mathcal{L}(a_n(\sum g_n - \mu_n) | \sum (X_i - \nu_n) = \nu) \xrightarrow{d} \mathcal{L}(Y|X = x)$. Let's see a concrete example of this:

Theorem 203

Let $N_1, \cdots, N_n$ be the counts of m balls in n boxes, and let $N_1', \cdots, N_n'$ be a separate count of independent m balls in n boxes. Let $D = \sum_{i=1}^n |N_i - N_i'|$. If $\frac{m}{n} \to \lambda > 0$, then $D \sim N(n\mu(\lambda), n\sigma^2(\lambda))$ is normally distributed.

Proof sketch. As usual, we randomize by realizing N_i, N_i' by independent Poisson(λ) random variables X_i, X_i', so that

$$\mathcal{L}(N_1, \cdots, N_n, N_1', \cdots, N_n') = \mathcal{L}\left(\vec{X}, \vec{X}' \middle| \sum_i X_i = \sum_i X_i' = m \right).$$

If we now define the variables

$$A_n = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} |X_i - X_i'| - \mu(\lambda)\right), \quad B_n = \frac{1}{\sqrt{n\lambda}}\left(\sum_{i=1}^{n} X_i - \lambda\right), \quad C_n = \frac{1}{\sqrt{n\lambda}}\left(\sum_{i=1}^{n} X_i' - \lambda\right),$$

then the conditioned law $\mathcal{L}(A_n|B_n = C_n = 0)$ is the same as the unconditioned law $\mathcal{L}\left(\frac{D-\mu(\lambda)}{\sqrt{n}}\right)$ for some random variable D. But $\begin{bmatrix} A_n \\ B_n \\ C_n \end{bmatrix} \to \begin{bmatrix} A \\ B \\ C \end{bmatrix}$ converges to a multivariate Gaussian with covariance matrix $\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & 0 \\ \rho & 0 & 1 \end{bmatrix}$, so in fact the limiting distribution $\mathcal{L}(A|B, C = 0)$ is $N(0, 1 - 2\rho^2)$, and unpacking the rest of our substitutions yields the result. □

Remark 204. *The proof above shows that in fact* $\begin{bmatrix} X \\ Y \end{bmatrix}$ *is infinitely divisible with law*

$$\mathbb{E}[e^{isX+itY}] = h(t)e^{as^2+bst+ct^2},$$

for $h(t)$ *an infinitely divisible characteristic function with no normal component. It turns out that if* $a_n \sum(g(X_i) - \mu_n)$ *converges to some random variable* X *with no normal component (say a Poisson), then the conditioning doesn't matter, but if* $\sum g(X_i)$ *converges to a normal random variable then the conditioning does.*

Le Cam used this method to study the distribution of functions of spacings (for example the sum of squares); in a setting like the conditional distribution $\sum w_i^4 | \sum w_i^2 = 1$, the conditioning does indeed matter. And the work that Holst did on this happened 50 years ago, so it could be "brought up to date" with new ideas!

Remark 205. *We can see Sourav Chatterjee's paper "A note about the uniform distribution on the intersection of a simplex and a sphere" for some further settings of the type where we study* $\sum x_i | \sum x_i^2 = 1$. *In particular, since* $(\sum x_i)^2 \le n \sum x_i^2$, *we know that* $|\sum x_i| \le \sqrt{n}$, *and we end up getting some interesting localization phenomena when we're close to that equality case.*