Introduction to Analytic Combinatorics

Zhenghua Lyu

Pivotal Inc.

Aug. 2019

Introduction to Analytic Combinatorics

Words from Knuth

"People who analyze algorithms have double happiness. First of all they experience the sheer beauty of elegant mathematical patterns that surround elegant computational procedures. Then they receive a practical payoff when their theories make it possible to get other jobs done more quickly and more economically."

Class and Generating Function

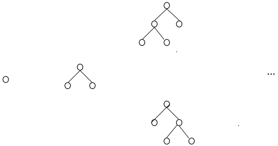
Definition

- Class is a set of elements with a size function
- Each element in Class has a size
- OGF : $\mathcal{A}$ is a class, $a \in \mathcal{A}$ is a element,
  $\mathcal{A}(z) = \sum_{a \in \mathcal{A}} z^{|a|} = \sum_{N \geq 0} A_N z^N, A_N = [z^n]\mathcal{A}(Z)$
- EGF : $\mathcal{A}$ is a class, $a \in \mathcal{A}$ is a element,
  $\mathcal{A}(z) = \sum_{a \in \mathcal{A}} \frac{z^{|a|}}{N!} = \sum_{N \geq 0} A_N \frac{z^N}{N!}, A_N = N![z^n]\mathcal{A}(Z)$
- A single function $\mathcal{A}(z)$ contains all the information we need of the class $\mathcal{A}$

Class and Generating Function

Example : unlabelled

- class $\mathcal{T}$ is all the binary trees, the size function is defined as $|t| = \#nodes\ in\ t$



- $\mathcal{T}(z) = z + z^3 + 2z^5 + \cdots = \frac{1-\sqrt{1-4z^2}}{2z}$

Class and Generating Function

Example : labelled

- class $\mathcal{P}$ is all permutations, the size function is defined as $|p| = length\ p$

empty	{1}	{1,2}	{1,2,3}	...
		{2,1}	{1,3,2}	
			{2,1,3}	
			{2,3,1}	
			{3,1,2}	
			{3,2,1}	

- $\mathcal{P}(z) = 1 + z + 2\frac{z^2}{2!} + 6\frac{z^3}{3!} + \cdots = \frac{1}{1-z}$

Properties of OGF

Add, Cartesian product, Sequence

- Add : $\mathcal{A}(z) + \mathcal{B}(z) \sim \mathcal{A} + \mathcal{B}$
- Cartesian product : $\mathcal{A}(z)\mathcal{B}(z) \sim \mathcal{A} \times \mathcal{B}$
- Sequence : $\frac{1}{1-\mathcal{A}(z)} \sim SEQ(\mathcal{A}), SEQ(\mathcal{A}) = \epsilon+\mathcal{A}+\mathcal{A}\times\mathcal{A}+\mathcal{A}\times\mathcal{A}\times\mathcal{A}+\ldots$

Properties of EGF

Add, Star product, Sequence, Set, Cycle

- Add : $\mathcal{A}(z) + \mathcal{B}(z) \sim \mathcal{A} + \mathcal{B}$
- Star product : $\mathcal{A}(z)\mathcal{B}(z) \sim \mathcal{A} \star \mathcal{B}$
- Sequence : $\frac{1}{1-\mathcal{A}(z)} \sim SEQ(\mathcal{A}), SEQ(\mathcal{A}) = \epsilon + \mathcal{A} + \mathcal{A} \star \mathcal{A} + \mathcal{A} \star \mathcal{A} \star \mathcal{A} + \dots$
- Set : $e^{\mathcal{A}(z)} \sim SET(\mathcal{A})$
- Cycle : $\ln \frac{1}{1-\mathcal{A}(z)} \sim CYC(\mathcal{A})$

Example of OGF

How many trees are there with N nodes ?

- Tree : a tree is one root with a forest
- $\mathcal{T} = \circ \times \mathcal{F} \implies \mathcal{T}(z) = z\mathcal{F}(z), T_{N-1} = F_N$
- Forest : a forest is several trees
- $\mathcal{F} = SEQ(\mathcal{T}) \implies \mathcal{F}(z) = \frac{1}{1-\mathcal{T}(z)}$
- $\mathcal{F}(z) = \frac{1-\sqrt{1-4z^2}}{2z}$
- Note forest has the same generating function as binary tree, this is not an accident, recall the algorithm that transferring between binary tree and forest

Example of EGF

100-prisoner problem

- Google's famous interview problem
- 100 prisoners numbered as $1, 2, \ldots, 100$, their numbers are randomly put into 100 boxes in one room
- the 100 boxes are also numbered as $1, 2, \ldots, 100$
- each time, only one prisoner can enter the room, and if he can find his number within opening 50 boxes, he succeeds
- prisoners are not allowed to communicate
- any prisoner fails, all 100 prisoners will be killed
- how to design a policy to let the 100 prisoners survive with a probability more that 0.3 ?

Example of EGF

100-prisoner problem using analytic combinatrics 1

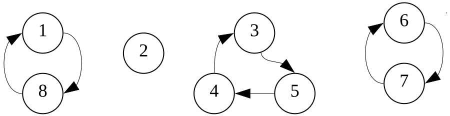
- Policy : each prisoner firstly opens the box with number equals his own number, then he will get a number in that box, and open the box with the previous number, continue this process...
- Each permutation corresponds to a set of cycles
- Numbers randomly putting in boxes is a permutation
- $Pr\{the\ 100\ prisoners\ survive\} = all\ cycles\ length \leq 50$

Example of EGF

Cycles and Permutations

- A permutation corresponds to a set of cycles

1 2 3 4 5 6 7 8
8 2 5 3 4 7 6 1



Example of EGF

100-prisoner problem using analytic combinatrics 2

- A permutation is a set of cycles
- $\mathcal{I}$ is the class of all permutations that contains no cycles whose length are larger than 50
- $\mathcal{I} = SET(CYC_1(Z)) \star SET(CYC_2(Z)) \star \cdots \star SET(CYC_{50}(Z))$
- $\mathcal{I}(z) = \prod_{i=1}^{50} e^{\frac{z^k}{k}} = e^{z+\frac{z^2}{2}+\cdots+\frac{z^{50}}{50}}$
- $[z^{100}]\mathcal{I}(z) \approx 0.31$ is just the probability

Stage 1 spilled tuples

Coupon Collector Model

- When generating 2 stage hash aggregation plan, we need to estimate how many tuples are streaming to stage 2(or how many tuples are spilled)
- On each segment, the hash table's capacity is hs, the data contains ng different groups, the size of each group is gs
- Data is in random order
- When the hash table is full, another tuple of a new group is coming, we stream all the data (hs tuples) in hash table to stage 2 and reset the hash table
- Q : How many tuples are streaming to stage 2 ?
- A : Asymptotically we can model this a Coupon Collector problem, $hs * gs / \ln(\frac{ng}{ng-hs})$ for $ng > hs$

Coupon Collector Model 1

Generating Function for Coupon Collector Model

- A coupon collector sequence is a M-slot with no empty slot
- Example of 5-slot : $\{7\}, \{1, 3\}, \{9\}, \{2, 4\}, \{5, 6, 8\}$
- $R_M = SEQ_M(SET_{>0}(Z))$
- EGF : $R_M(z) = (e^z - 1)^M$
- The remaining analysis involves complex mathematical process
- Average Wait time for first coupon collector sequence is MH_M
- H_M is harmonic series : $\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{M}$

Coupon Collector Model 2

Inclusion-exclusion principle

- $Average\_wait\_time = \sum_{N \geq 0} Pr\{not\ done\ for\ N\ steps\}$
- $Pr\{not\ done\ for\ N\ steps\} = Pr\{\sum_{i=1}^{N} Event_i\}$
- $Event_i$: missing slot i
- $Pr\{\sum_{i=1}^{N} Event_i\} = \sum_{i=1}^{N} (-1)^{i+1} \binom{M}{i} (\frac{M-i}{M})^N$
- $Average\_wait\_time = MH_M$

Result of Monte Carlo Simulation

- It is hard to analyze the performance of the model
- C++ code to simulate
- The model works quite well, for large number of groups, error rate is much less than 1%

Ramanujan

Genius of Genius

$$\frac{1}{\pi} = \frac{1}{8}\sum_{m=0}^{\infty}(20m+3)\frac{(-1)^m(4m)!}{(4\sqrt{2})^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{\sqrt{3}}{16}\sum_{m=0}^{\infty}(28m+3)\frac{(-1)^m(4m)!}{(64\sqrt{3})^{2m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{1}{72}\sum_{m=0}^{\infty}(260m+23)\frac{(-1)^m(4m)!}{(12\sqrt{2})^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{1}{18\sqrt{11}}\sum_{m=0}^{\infty}(280m+19)\frac{(4m)!}{(12\sqrt{11})^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{2}{84^2}\sum_{m=0}^{\infty}(21460m+1123)\frac{(-1)^m(4m)!}{(84\sqrt{2})^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{1}{2\sqrt{3}}\sum_{m=0}^{\infty}(8m+1)\frac{(4m)!}{(4\sqrt{3})^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9}\sum_{m=0}^{\infty}(10m+1)\frac{(4m)!}{12^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{3\sqrt{3}}{49}\sum_{m=0}^{\infty}(40m+3)\frac{(4m)!}{28^{4m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{\sqrt{5}}{288}\sum_{m=0}^{\infty}(644m+41)\frac{(-1)^m(4m)!}{(1152\sqrt{5})^{2m}(m!)^4}$$

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{99^2}\sum_{m=0}^{\infty}(26390m+1103)\frac{(4m)!}{396^{4m}(m!)^4}$$

FIGURE – Some formulas discovered by Ramanujan

The Man Who Knew Infinity

1729

FIGURE – The Man Who Knew Infinity

Integer Partition

Ramanujan's Formula

- $P(n)$: the number of ways to partition a number N
- $\mathcal{P} = MSET(\mathcal{I}) \implies \mathcal{P}(z) = e^{\mathcal{I}(z)+\frac{1}{2}\mathcal{I}(z^2)+\ldots}$
- $\mathcal{P} = \prod_{m=1}^{\infty} \frac{1}{1-z^m} = 1 + z + 2z^2 + 3z^3 + 5z^4 \ldots$
- Ramanujan says $\mathcal{P}_N \sim \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}}$

Analysis of Algorithms

Quick Sort

- $C_N = N + 1 + \frac{1}{N}\sum_{1\le j\le N}(C_{j-1} + C_{N-j})$
- $\frac{dC(z)}{dz} = \frac{2}{(1-z)^3} + 2\frac{C(z)}{1-z}$
- $C(z) = \frac{2}{(1-z)^2}\ln\frac{1}{1-z}$
- $C_N = 2(N+1)(H_{N+1} - 1)$

References

- Sedgewick, Robert, and Philippe Flajolet. An introduction to the analysis of algorithms. Pearson Education India, 2013.
- Flajolet, Philippe, and Robert Sedgewick. Analytic combinatorics. cambridge University press, 2009.
- Ronald L. Graham, Donald E. Knuth and Oren Patashnik, Concrete Mathematics, 2nd Edition.

Q & A

Thanks!
Q & A