SUMMARY: LAGRANGE INVERSION AND RANDOM FORESTS

GÁBOR PETE

1. The Lagrange inversion formula

If $f(z) = \sum_{n\geq 1} a_n z^n$ and $f^k(z) = \sum_{n\geq k} a_n^{(k)} z^n$, then by forming the upper-triangular matrix $A_f := \left(a_n^{(k)}\right)_{k,n\geq 1}$ we have established an endomorphism $f \mapsto A_f$ from the group of formal power series with compositional inverse, into the group of invertible upper-triangular matrices. The coefficients of $f^{[-1]}(z) = \sum_{n\geq 1} b_n z^n$ form the first row of the inverse matrix $A_f^{-1} = A_{f^{[-1]}} = \left(b_n^{(k)}\right)_{k,n\geq 1}$. So our task is to invert an infinite-dimensional matrix! Fortunately, the answer can be derived in a completely different way. The simplest version of the Lagrange inversion formula, also called the **Schur-Jabotinski formula**, says

$$b_n^{(k)} = \frac{k}{n} a_{-k}^{(-n)}. \tag{1}$$

In the book I would point out the connection to **Jabotinski's matrix interpretation** of the composition of exponential generating functions, and to the **Faà di Bruno formula.**

There are **multivariate Lagrange expansion** formulae, as well, the first of which was formulated by Jacobi [Jac30], and was proved in full generality by Good [Goo60]. For the different versions of single and multivariate Lagrange formulae, including some different **Jacobi formulae**, a good reference is [Ges87].

The gap that I found anno in [Pit98, Thm 1.4] fortunately doesn't appear in the book. That was a difficult bijection between two sets that are not convenient to compare directly.

If the offspring distribution X_i happens to be infinitely divisible, there is a nice continuous time queuing-like process to build up the corresponding GW-tree. Take the Lévy-process Y_t with $Y_1 \stackrel{d}{=} X_i$, and look at $Z_t = Y_t - t$ from $t = 0$ till $t = T_{-1}$, the first passage time to -1, a positive integer. Let us suppose that Y_t has jumps of size 1, so the number of jumps of $\{Y_t : 0 \leq t \leq 1\}$ has also a distribution X_i; this corresponds to the assumption that only one customer arrives at one time, so we can decide the order of them. Let $V_t = \inf\{Z_s : 0 \leq s \leq t\}$. The length of time intervals inside $[0, T_{-1}]$ where $V_t = Z_t$ add up to 1, and the number of time intervals where V_t is constant has distribution X_i. Let the interval $[0, T_{-1}]$ be the root of our GW-tree, and the constant intervals of V_t be the children of the root. Now we can replay the procedure for each of these children, etc. The size of the resulting GW-tree will be T_{-1}.

Date: November 17, 2003— Remarks on **Jim Pitman**'s book in preparation, *Combinatorial Stochastic Processes*.

2. Lagrangian distributions

The probability distributions on $\{0, 1, 2, \dots, \infty\}$ which arise as the total progeny distribution of a GW forest with Z_0 distributed according to any given discrete random variable, are called **Lagrangian distributions**. They were defined the first time in [CS72], but the GW description was given only by [Goo75]. The former paper notices that the composition of two Lagrangian distributions is also Lagrangian, which is obvious from the GW description. Some examples of Lagrangian distributions: geometric, Borel-Tanner (see in next section), negative binomial.

A nice interesting paper is by Viskov [Vis00]. He gives an algebraic proof of the Lagrange inversion formulae, with the representation theory of the Heisenberg-Weyl algebra, as the underlying idea. He deduces a new, exponential version of the inversion formula, which allows him to prove that if $h(z)$ is a basic Lagrangian distribution, i.e. the total progeny of a single GW-tree with offspring p.g.f. $g(0) \neq 0$, then it is **infinitely divisible**, with a possible positive mass at infinity. In fact, for $G(z) = h(z)/z$, $G^\lambda(z)$ is the generating function of a compound Poisson process $Y_\lambda,\ \lambda > 0$,

$$P(Y_\lambda = m) = \frac{\lambda}{m!(\lambda+m)} \frac{d^m}{dx^m} \Big[g^{\lambda+m}(x) \Big]_{x=0}. \tag{2}$$

The Lévy-Khintchin formula is

$$G^\lambda(z) = \exp\left\{ \lambda \left[\log g(0) + \sum_{n=1}^{\infty} \frac{z^n}{n!n} \frac{d^n}{dx^n} \Big[g^n(x) \Big]_{x=0} \right] \right\}.$$

(Here I have **a problem:** if there is no mass at infinity, i.e. $g'(1) \leq 1$, then do the rates of the compound Poisson process sum up to $-\log g(0)$? They don't seem like that....)

Note that for $\lambda = k \in \mathbb{Z}^+$, $G^k(z)$ has a clear probabilistic meaning, and $Y_k + k$ coincides with the first passage time T_{-k} of the Kemperman setting. But what about $\lambda = 1/2$, say? Viskov also considers Bernoulli random walks where the jump times are Poisson randomized.

Limit theorems by [PS77]: what is the limiting distribution of the total progeny (correctly normalized) as the expected values ν of Z_0 and μ of the offspring distribution approach certain values, e.g. $\mu \to 1$, $\nu \to \infty$, $\nu(1-\mu) \to c \in (0, \infty)$. Takács proves limit theorems for the height and diameter of GW-trees with offspring distributions arising naturally from the queuing interpretation [Tak93].

3. Simply generated forests

A **labeled forest** means a forest of vertex-labeled rooted trees, and a **plane forest** consists of unlabeled rooted trees with an ordering of each level. From the set $\mathbf{F}^{[n]}_{k,n}$ of labeled forests of n vertices and k trees, there is a natural map $\mathbf{f} \mapsto \mathbf{f}^\circ$ onto the set $\mathbf{F}^\circ_{k,n}$ of plane forests of the same size: order the vertices on each level by the order of their labels.

A **simply generated forest** is a probability measure on $\mathbf{F}^\circ_{k,n}$, given by conditioning a GW forest with k initial individuals and offspring distribution p.g.f. $F(z)$ on having total offspring size n. Note that the uniform distribution $\mathcal{F}^{[n]}_{k,n}$ is not such a thing, but

$$\mathcal{F}^{[n]}_{k,n} \stackrel{d}{=} (\mathcal{P}^*_{k,\mu} \,|\, \#\mathcal{P}_{k,\mu} = n), \tag{3}$$

where $\mathcal{P}^*_{k,\mu}$ is the Poisson GW forest labeled by a uniform random permutation of the vertex set. To prove this, one needs the following formula for the **Borel-Tanner distribution**:

$$P(\#\mathcal{P}_{k,\mu} = n) = \frac{k}{n}\frac{(\mu n)^{n-k}}{(n-k)!}\, e^{-\mu n}\ \ (n = k, k+1, \ldots), \tag{4}$$

which can be deduced from the **Otter-Dwass formula**

$$P(\#\mathrm{GW}_k(F(z)) = n) = \frac{k}{n} P(S_n(F(z)) = n-k), \tag{5}$$

which is basically equivalent to Kemperman's formula, the cycle lemma, and to the Lagrange inversion formula, see [Pit98] and [Wen75].

Similar conditional counting identities appear in [SMS94], they are quite interesting, and might also be good for exercises.

The vector of component sizes $\nu_1, \ldots, \nu_k$ of a simply generated forest is an example of a **generalized allocation scheme**. This means that there exist i.i.d. variables $\xi_1, \ldots, \xi_k$ such that $P(\nu_i = n_i,\ i = 1, \ldots, k) = P(\xi_i = n_i,\ i = 1, \ldots, k \,|\, \sum_i \xi_i = n)$ for arbitrary values $\sum_i n_i = n$. The simplest example is the **classical allocation scheme**, where we want to distribute n balls into k cells: if the number of balls in cell i is ν_i, then $\xi_i \sim$Poisson(λ) will work for arbitrary $\lambda > 0$.

This **λ-invariance** of Poisson(λ) inspires a nice more general fact about simply generated forests: if we have a GW process with some offspring distribution p_i and p.g.f. $F(z)$, then for any $0 < \lambda \le 1$ the GW-process with offspring distribution $p_i(\lambda) = \lambda^i p_i / F(\lambda)$, p.g.f. $F_\lambda(z) = F(\lambda z)/F(\lambda)$, defines the same simply generated forest. (The reason is that conditioned on the total size to be n, if the number of vertices with i children is c_i, then $\sum_i c_i = \sum_i i c_i = n$.) In particular, if p_i is Poisson(1), then $F(z) = e^{z-1}$ and $p_i(\lambda)$ is Poisson(λ), so the classical allocation scheme shows that Poisson(λ) GW forests have an even finer λ-invariance than arbitrary GW forests. However, the general λ-invariance is still important in many ways.

The expected offspring size corresponding to $p_i(\lambda)$ is $m_\lambda = \lambda F'(\lambda)/F(\lambda)$, and the expected tree size is $1/(1-m_\lambda)$. Thus choosing λ as $m_\lambda = (n-k)/n$ will make the expected total forest size (without conditioning) exactly n, and then we can hope in transporting some unconditional results more easily into the conditional world. Indeed, [Pav00] establishes asymptotics for component sizes by calculating separately the three factors in the identity

$$P(\max_i \nu_i \le r) = (1 - P(\xi_1 > r))^k \frac{P(\sum_i \xi_i = n \,|\, \xi_i \le r,\ i = 1, \ldots, k)}{P(\sum_i \xi_i = n)}, \tag{6}$$

using and proving conditional local limit theorems with the above choice of λ.

Another cute fact about Poisson GW trees from [AS92]: if $\lambda < 1 < \mu$ is **a conjugate pair** in the sense that $\lambda e^\lambda = \mu e^{-\mu}$, then for the corresponding Poisson GW trees $\mathcal{P}_{1,\lambda} \overset{d}{=} (\mathcal{P}_{1,\mu} \,|\, \#\mathcal{P}_{1,\mu} < \infty)$. **More generally**, as I have just observed, given any offspring distribution p.g.f. $F(z)$ with survival probability $q < 1$, there is exactly one $\lambda \in (0,1)$ such that

$$\mathrm{GW}_1(F_\lambda(z))) \overset{d}{=} \big(\mathrm{GW}_1(F(z)) \,\big|\, \#\mathrm{GW}_1(F(z)) < \infty\big), \tag{7}$$

and this value is $\lambda = q$.

I think that the λ-invariance and this general conjugation together could form a good **exercise** before Exercise 1 on page 103.

Also, I would explicitely remark that this Exercise 1 implies that if $d < 1 < c$ are conjugates in the Poisson sense, then deleting the giant component from $\mathcal{G}(n, c/n)$ results in a random graph which is basically $\mathcal{G}(n', d/n')$, where n', the number of vertices not in the giant component, satisfies $n' \sim nq$, where $q = d/c$ is the extinction probability.

4. Giant components

Using the **multiplicative coalescent**, the book explains the emergence of the giant component in $\mathcal{G}(n, p(n))$ at $p(n) \sim 1/n$, or equivalently, in $\mathcal{G}(n, m(n))$ at $m(n) \sim n/2$, see [Ald97]. The last paragraph in the previous section points out the **self-similarity** in the dynamical structure. Very similar, but strangely different phenomena can be found in the following two random forest models: **1.** $\mathcal{F}^{[n]}_{k,n}$, or in general, simply generated forests; **2.** The uniform **labeled unrooted forest** $\mathcal{G}^{[n]}_{k,n}$, which is clearly closer to the unrooted Erdős-Rényi model $\mathcal{G}(n, m(n))$ than the rooted model $\mathcal{F}^{[n]}_{k,n}$.

For these two forest models the emergence of the giant component happen in two different regimes: at $m(n) \sim n/2$ (i.e. $k(n) \sim n/2$) for $\mathcal{G}^{[n]}_{k,n}$, and at $k(n) \sim \sqrt{n}$ for $\mathcal{F}^{[n]}_{k,n}$. However, independently of this difference, the orders of magnitude of the size of the largest component are the same for the two models all along the evolution: the difference is in the second largest component. Moreover, the behaviour of this second largest component is not completely the same in $\mathcal{G}^{[n]}_{k,n}$ and in $\mathcal{G}(n, m(n))$: by the end of the critical regime, the second largest component of the forest doesn't drop down to $\log n$, but stays $n^{2/3}$.

The simply generated forests are nicer models, and the $\mathcal{F}^{[n]}_{k,n}$ is actually intimately connected to the **standard additive coalescent**, see [AP98]. The book contains a lot of things about the SAC, but doesn't point out that all the results of [Pav00] and [Che98] follow from the SAC. To summarize briefly:

According to (4), if $\xi \sim \#\mathcal{P}_{1,1}$ is the size of a Poisson(1) GW tree, then $P(\xi = j) = e^{-1} j^{j-1}/j!$, and by Stirling's formula

$$P(\xi = j) \sim (2\pi)^{-1/2} j^{-3/2}.$$

So if Z_1 has the 1/2-stable density $g(x) = (2\pi)^{-1/2} x^{-3/2} \exp(-1/2x)$ for $x \geq 0$, then for i.i.d. copies of ξ we have

$$\frac{1}{n} \sum_{i=1}^{n^{1/2}} \xi_i \Longrightarrow Z_1. \tag{8}$$

Moreover, a local limit theorem holds. So it's not surprising that for $k(n) \sim cn^{1/2}$ the largest and the second largest etc. component sizes, when normalized by n, converge to some non-degenerate distribution involving the stable distribution $g(x)$ in some way. Furthermore, in the regime $n(k)/k^2 \to \infty$, where the conditional total size n is much larger than the unconditional expected size k^2, a single giant component emerges, such that the remaining components already behave exactly like an unconditional GW forest with $k - 1$ trees.

The complete proofs of [Pav00] and [Che98] are based on (6). For a more conceptual proof one should generalize the SAC-results from Poisson GW forests to arbitrary offspring distributions. The general conjugacy (7) above might be the key to self-similarity. **Am I right here?** Is this the same as p-forests and general additive coalescents?

Now we turn to the $\mathcal{G}_{k,n}^{[n]}$ model. First recall **Rényi's formula** on the number of labeled unrooted forests:

$$g(n,k) = \#\mathbf{G}_{k,n}^{[n]} = \frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}(-\frac{1}{2}^{i})(k+i)n^{n-k-i-1}(n)_{k+i}. \tag{9}$$

We will write $f(n,m) = g(n,n-m)$ to agree with our references. This formula can easily be deduced from the exponential generating function

$$\mathcal{T}(z) = \sum_{j\geq 1}\frac{j^{j-2}z^j}{j!} = T(z) - T^2(z)/2 \tag{10}$$

of unrooted labeled trees, where $T(z) = \sum_{j\geq 1}\frac{j^{j-1}z^j}{j!} = z\exp(T(z))$ is the exponential generating function of rooted labeled trees, a very handy formal power series. Clearly,

$$f(n,m) = \frac{n!}{(n-m)!}[z^n]\mathcal{T}^{n-m}(z). \tag{11}$$

The asymptotic behaviour of $f(n,m)$ was studied by Britikov [Bri88], and actually these results form the main ingredient in the study [ŁP92] of the emergence of the giant component in $\mathcal{G}_{k,n}^{[n]}$.

The key observation in obtaining these asymptotics is that

$$[z^n]\mathcal{T}^{n-m}(z) = \frac{\mathcal{T}^{n-m}(\zeta)}{\zeta^n}P(\sum_{i=1}^{n-m}Y_i = n), \tag{12}$$

where $\zeta \in (0, e^{-1})$ is fixed, and the Y_i's are i.i.d. variables with p.g.f. $E(z^{Y_i}) = \mathcal{T}(\zeta z)/\mathcal{T}(\zeta)$. Now ζ can be set freely to obtain $EY_i = n/(n-m)$, and then can use local limit theorems to estimate the probability factor in (12) — an idea we already saw above for simply generated forests.

Then one proves that these i.i.d. random variables Y_i belong to the domain of attraction of a stable distribution with parameter 2 (normal distribution) in the subcritical case $s^3/n^2 \to -\infty$, where $s = 2m-n$, and to the domain of a 3/2-stable in the critical case $|s|^3/n^2 < C$, and also in the supercritical case $s^3/n^2 \to \infty$.

Once we have these asymptotics, it is good to notice that if $X_{n,m}(j_1,j_2)$ denotes the number of components of size in $[j_1,j_2]$, then

$$E(X_{n,m}(j_1,j_2)) = \sum_{r=j_1}^{j_2}\binom{n}{r}r^{r-2}\frac{f(n-r,m-r+1)}{f(n,m)}, \tag{13}$$

and a similar expression holds for the j-th factorial moments $E_jX_{n,m}(j_1,j_2)$. Then the main steps are the following:

1. In the entire subcritical regime the factorial moments of $X_{n,m}$ can be well approximated by the factorial moments of the corresponding $Y_{n,m}$ variables for $\mathcal{G}(n,m)$, and so we are done with this regime.

2. If $sn^{-2/3} \to \alpha$, then the size of the largest component is $O_p(n^{2/3})$. This can be proved by the estimate

$$E(X_{n,m}(\omega(n)n^{2/3}, n)) \leq c \int_{\omega(n)}^{\infty} x^{-2} \exp(-x^3) dx \to 0, \tag{14}$$

when $\omega(n) \to \infty$.

3. In the same critical regime as in step 2, for $0 < d < D < \infty$ arbitrary constants, a formula for $E_j(X_{n,m}(dn^{2/3}, Dn^{2/3}))$ can be achieved, which shows that $X_{n,m}(dn^{2/3}, n) \Longrightarrow X(d)$ where

$$E_j(X(d)) = \frac{1}{2\pi p(\alpha)} \int_{-\infty}^{\infty} e^{-it\alpha} \phi(t) \left(\frac{1}{\sqrt{2\pi}} \int_d^{\infty} \frac{e^{itx}}{x^{5/2}}\, dx \right)^j dt, \tag{15}$$

where $\phi(t)$ is the characteristic function of a 3/2-stable distribution $p(x)$, appearing already in the asymptotics for $f(n, m)$.

4. If $s^3/n \to \infty$ but $n - s \to \infty$, then for every constant d,

$$\frac{|\text{largest component of } \mathcal{G}_{k,n}^{[n]}| - s}{(n-s)^{2/3}} \Longrightarrow p(-x). \tag{16}$$

Here the key is that we can estimate the number of forests with largest component of size $r \in [s - D(n-s)^{2/3}, s - d(n-s)^{2/3}]$, by building them starting with a component of size r in one from $\binom{n}{r} r^{r-2}$ possible ways, and then take a random forest with $m' = m - r + 1$ edges on the remaining $n' = n - r$ vertices — by step 2, the largest component here will be $O((n')^{2/3}) = o(r)$.

5. The previous argument also shows that deleting the largest component from a supercritical forest we arrive at a critical forest, so the smaller components can be obtained by step 3.

References

[Ald97] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854, 1997.

[AP98] D.J. Aldous and J. Pitman. The standard additive coalescent. *Ann. Probab.*, 26:1703–1726, 1998.

[AS92] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley, 1992.

[Bri88] V. E. Britikov. Asymptotics of the number of forests made up of nonrooted trees. *Mat. Zametki*, 43(5):672–684, 703, 1988.

[Che98] I. A. Cheplyukova. The emergence of a giant tree in a random forest. *Diskret. Mat.*, 10(1):111–126, 1998.

[CS72] P. C. Consul and L. R. Shenton. Use of Lagrange expansion for generating discrete generalized probability distributions. *SIAM J. Appl. Math.*, 23:239–248, 1972.

[Ges87] I. Gessel. A combinatorial proof of the multivariable Lagrange inversion formula. *J. Combinatorial Theory A*, 45:178–195, 1987.

[Goo60] I. J. Good. Generalizations to several variables of Lagrange's expansion, with applications to stochastic processes. *Proc. Cambridge Philos. Soc.*, 56:367–380, 1960.

[Goo75] I.J. Good. The Lagrange distributions and branching processes. *SIAM Journal on Applied Mathematics*, 28:270–275, 1975.

[Jac30] C. G. J. Jacobi. De resolutione aequationum per series infinitas. *J. Reine Angew. Math.*, 6:257–286, 1830.

[ŁP92] T. Łuczak and B. Pittel. Components of random forests. *Combin. Probab. Comput.*, 1(1):35–52, 1992.

[Pav00] Yu. L. Pavlov. *Random forests*. VSP, Utrecht, 2000.

[Pit98] J. Pitman. Enumerations of trees and forests related to branching processes and random walks. In D. Aldous and J. Propp, editors, *Microsurveys in Discrete Probability*, number 41 in DIMACS Ser. Discrete Math. Theoret. Comp. Sci, pages 163–180, Providence RI, 1998. Amer. Math. Soc.

[PS77] A. G. Pakes and T. P. Speed. Lagrange distributions and their limit theorems. *SIAM Journal on Applied Mathematics*, 32:745–754, 1977.

[SMS94] M. Sibuya, N. Miyawaki, and U. Sumita. Aspects of Lagrangian probability distributions. *Studies in Applied Probability. Essays in Honour of Lajos Takács (J. Appl. Probab.)*, 31A:185–197, 1994.

[Tak93] L. Takács. Limit distributions for queues and random rooted trees. *J. Applied Mathematics and Stochastic Analysis*, 6:189–216, 1993.

[Vis00] O. V. Viskov. A random walk with an upper-continuous component, and the Lagrange inversion formula. *Teor. Veroyatnost. i Primenen.*, 45(1):166–175, 2000.

[Wen75] J.G. Wendel. Left continuous random walk and the Lagrange expansion. *Amer. Math. Monthly*, 82:494–498, 1975.

Department of Statistics, UC Berkeley
E-mail address: gabor@stat.berkeley.edu