Exercises for the course *Data Stream Analysis: a (new) triumph for Analytic Combinatorics*

ALEA in Europe Workshop, Vienna (Austria)

Conrado Martínez

October 2017

1. The (signless) Stirling numbers of the first kind $\left[{n \atop k}\right]$ satisfy the following recurrence

$$\left[{n \atop k}\right] = (n-1)\left[{n-1 \atop k}\right] + \left[{n-1 \atop k-1}\right], \qquad n > 0.$$

By convention, we take $\left[{0 \atop 0}\right] = 1$ and $\left[{n \atop k}\right] = 0$ whenever $n < k$.

   (a) Give a combinatorial argument to show that $\left[{n \atop k}\right]$ is the number of permutations of size n that contain exactly k cycles.

   (b) From the recurrence for $\left[{n \atop k}\right]$, compute a functional equation satisfied by the bivariate generating function

$$S(z,u) = \sum_{n\geq 0}\sum_{k\geq 0} \frac{\left[{n \atop k}\right]}{n!} z^n u^k$$

   Hint: $S(z,u)$ satisfies a linear partial differential equation.

2. The class of permutations $\mathcal{P}$ can be specified as

$$\mathcal{P} = \textsc{Set}(\textsc{Cycle}(Z)),$$

that is, each permutation can be seen as a set of labelled cycles.

   (a) Use the symbolic method to find the bivariate generating function

$$C(z,u) = \sum_{\sigma\in\mathcal{P}} \frac{z^{|\sigma|}}{|\sigma|!} u^{c(\sigma)},$$

   where $c(\sigma)$ is the number of cycles in the permutation σ.

   (b) Show that the explicit form for $C(z,u)$ satisfies the functional equation obtained in the previous exercise for $S(z,u)$.

3. Perform the detailed computation of $\mathrm{Var}[R_n]$, with R_n the number of k-records in a random permutation of size n, starting from the explicit form for $\Phi(z, u)$.

4. Since $\mathrm{E}[R_n] = k \ln(n/k) + O(1)$ we could use

$$Z' = k \exp(R/k) \cdot \varphi,$$

as our cardinality estimator, with φ a correcting factor to make sure that Z' is an asymptotically unbiased estimator of n, that is, $\mathrm{E}[Z'] = n + o(n)$. Find a closed form for the correcting factor φ.

5. After execution of RECORDINALITY (if $k \leq n$) the table T contains a random sample of k distinct eleemnts. Why? Suppose that we modify the algorithm as follows: for each incoming elemnt s with hash value x:

   - If s is in the table or x is smaller than the minimum hash value in T then discard s.
   - If x is larger than the k-th largest hash value in T, add s to T. No element from T is removed.
   - If x is smaller than the k-th largest hash value in T but larger than the minimum value in T, add s to T and remove the element with minimum hash value.

   (a) What does T contain after execution of the algorithm?

   (b) The size of T is now given by a random variable. What is the expected size?

   (c) Bonus problem. Each time an element "kicks out" some element from T we say that there is a *replacement*. What is the expected number of replacements $\mathrm{E}[f_n]$ for a random permutation of size n? Consider first the simpler case when $k = 1$. Hint: Let Y_i be the indicator random variable for the event that the ith incoming element is a replacement. Compute the probability that $Y_i = 1$ conditioned on $R_{i-1} = j$ (i.e., the number of records seen so far is j). Uncondition and obtain $\mathrm{E}[f_n]$ by linearity of expectation.