计算多个序列碰撞的概率

问题 1：计算两个序列碰撞的概率

$\{a\}$ 是一个长度为 n 的数列，$a_1, a_2, ..., a_n$ 是 $1, 2, ..., n$ 的某个排列。$\{b\}$ 也是一个长度为 n 的数列，$b_1, b_2, ..., b_n$ 是 $1, 2, ..., n, ..., m$ 的某个排列的前 n 项 $(n \le m)$，($1, 2, ..., n$ 的某个排列和 $1, 2, ..., n, ..., m$ 的某个排列都是从所有排列中等概率选取的)。若存在某个正整数 k 使得 $a_k = b_k$，则称 $\{a\}$ 与 $\{b\}$ 碰撞了一次。求 $\{a\}$ 与 $\{b\}$ 碰撞 i 次的概率 $p_{n,m,2}(i)$ $(0 \le i \le n)$。

解

先计算 $\{a\}$ 与 $\{b\}$ 碰撞 0 次的概率，并且假设 $n \le m \le 2n$，$m > 2n$ 的情况可以很简单地从 $n \le m \le 2n$ 的结果推出。

显然可以固定数列 $\{a\}$，设 $a_k = k, 1 \le k \le n$。设 S_n 为 $m = n$ 时，$\{a\}$ 与 $\{b\}$ 碰撞次数为 0 的个数。设数列 $\{c\} = (1, 2, ..., n, ..., m)$，把 $\{c\}$ 分为两部分 $\{c_{left}\}$ 和 $\{c_{right}\}$，$\{c_{left}\} = (1, 2, ..., n)$,$\{c_{right}\} = (n+1, n+2, ..., m)$。设 $\{b'\}$ 是 $\{c\}$ 的某个排列的前 n 项，且 $\{b'\}$ 中包含 $\{c_{right}\}$ 中的所有项，设 T_m^n 是 $\{a\}$ 与 $\{b'\}$ 碰撞次数为 0 的的个数，于是有 $S_n = T_n^n$ (如何用 T_m^n 计算概率)。下面先计算 S_n，再计算 T_m^n。

S_n 是 $m = n$ 时，$\{a\}$ 与 $\{b\}$ 碰撞次数为 0 的个数，$\{b\}$ 可以由如下方式选取：

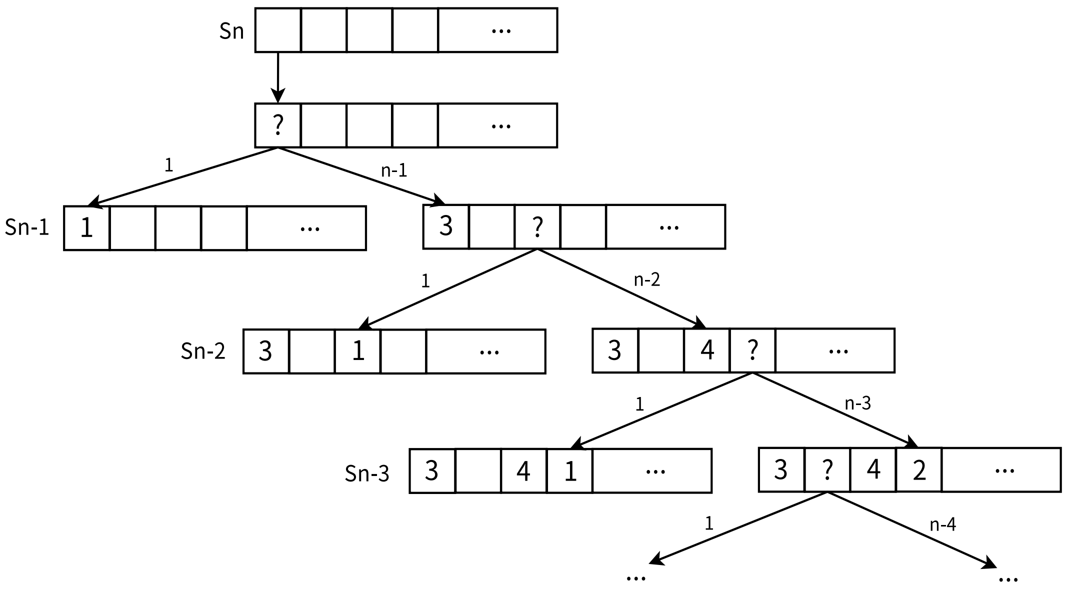

图 1: S_n 的递归计算

先选择 b_1，有 $n-1$ 种可能，假设选择了 3, 然后选择 b_3，若 $b_3 = 1$，则剩下 $n-2$ 个数的无碰撞排列个数为 S_{n-2}; 否则 b_3 有 $n-2$ 中可能，假设选择了 4, 然后选择 b_4，若 $b_4 = 1$，则剩下 $n-3$ 个数的无碰撞排列个数为 S_{n-3}; 否则 b_4 有 $n-3$ 中可能，... 。以此类推。

$$S_n = (n-1)(S_{n-2} + (n-2)(S_{n-3} + (n-3)(S_{n-4} + \cdots$$

$$S_{n-1} = (n-2)(S_{n-3} + (n-3)(S_{n-4} + \cdots$$

于是有：

$$S_n = (n-1)(S_{n-1} + S_{n-2}) \qquad S_2 = 1,\ S_1 = 0 \tag{1}$$

T_m^n 是 $\{a\}$ 与 $\{b^{'}\}$ 碰撞次数为 0 的的个数。设 $\{c^{'}\}$ 为 $\{c\}$ 的某个排列，且 $\{c^{'}\}$ 与 $\{c\}$ 的碰撞次数为 0。$\{c^{'}\}$ 可分为两部分 $\{c^{'}_{left}\}$ 和 $\{c^{'}_{right}\}$。S_m 可分为 $m-n+1$ 个部分：

$\{c^{'}\}$ 的个数, $\{c^{'}_{left}\}$ 包含 $\{c_{right}\}$ 中的某 0 项

$\{c^{'}\}$ 的个数, $\{c^{'}_{left}\}$ 包含 $\{c_{right}\}$ 中的某 1 项

$\{c^{'}\}$ 的个数, $\{c^{'}_{left}\}$ 包含 $\{c_{right}\}$ 中的某 2 项

$\vdots$

$\{c^{'}\}$ 的个数, $\{c^{'}_{left}\}$ 包含 $\{c_{right}\}$ 中的某 $m-n$ 项

设 R_k 为 $\{c^{'}\}$ 的个数, $\{c^{'}_{left}\}$ 包含 $\{c_{right}\}$ 中的某 k 项 $(0 \le k \le m-n)$，于是 $S_m = \sum\limits_{k=0}^{m-n} R_k$。

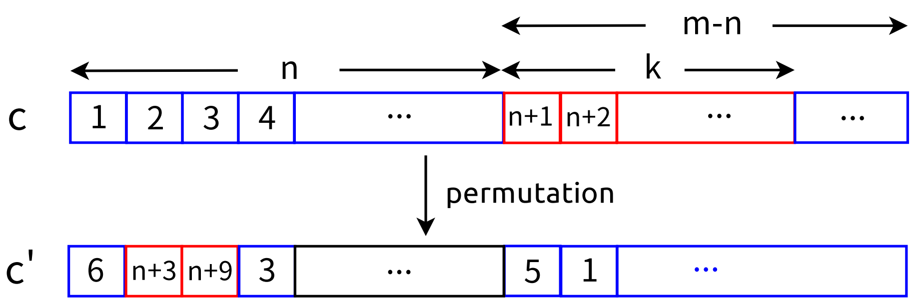

图 2: c 的某个排列

根据 T_m^n 的定义，$R_k = C_{m-n}^k T_{n+k}^n \dfrac{T_{m-n+k}^{m-n}}{C_{m-n}^k} = T_{n+k}^n T_{m-n+k}^{m-n}$。其中 $C_{m-n}^k T_{n+k}^n$ 表示从 $\{c_{right}\}$ 中选出的 k 项放入 $\{c^{'}_{left}\}$ 能产生多少个与 $\{c_{left}\}$ 碰撞 0 次的 $\{c^{'}_{left}\}$; **在此基础上**，$\{c_{left}\}$ 中确定的 k 项会被放入 $\{c^{'}_{right}\}$，因此 $\dfrac{T_{m-n+k}^{m-n}}{C_{m-n}^k}$ 的分子中有 T_{m-n+k}^{m-n} 而没有 C_{m-n}^k。分母中的 C_{m-n}^k 表示 $C_{m-n}^k T_{n+k}^n T_{m-n+k}^{m-n}$ 计算重复了，因为每一个被放入 $\{c^{'}_{right}\}$ 的 k 项都对应着从 $\{c_{right}\}$ 中选出的 k 项的 C_{m-n}^k 个组合。于是：

$$T_m^m = S_m = \sum_{k=0}^{m-n} R_k = \sum_{k=0}^{m-n} T_{n+k}^n T_{m-n+k}^{m-n} \qquad T_0^0 = 1 \tag{2}$$

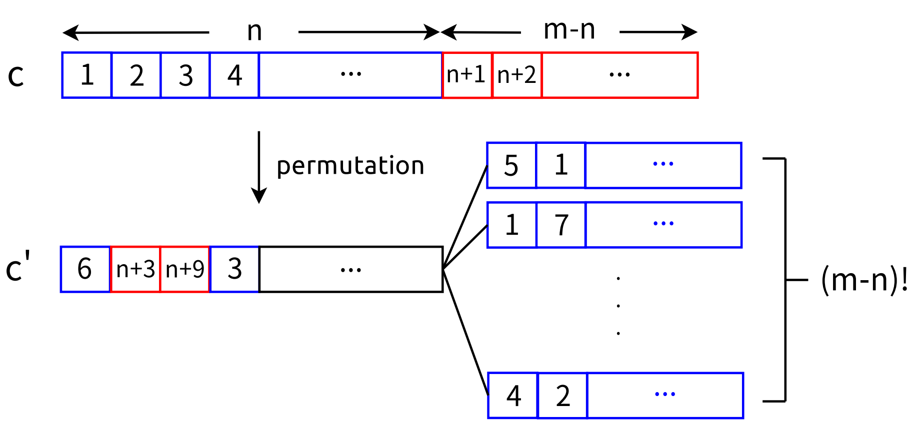

图 3: R_{m-n} 重复计算的部分

因为 R_k 表示的是 $\{c_{left}\}$ 与 $\{c_{left}^{'}\}$ 碰撞次数为 0 且 $\{c_{right}\}$ 与 $\{c_{right}^{'}\}$ 碰撞次数为 0 的个数，$\{c_{left}^{'}\}$ 包含 $\{c_{right}\}$ 中的某 $m-n$ 项，而需要计算的只是 $\{c_{left}\}$ 与 $\{c_{left}^{'}\}$ 碰撞次数为 0 的个数。如图 3所示，每一个被放入 $\{c_{right}^{'}\}$ 的 $m-n$ 项不论如何排列都不会与 $\{c_{right}\}$ 碰撞，而这 $(m-n)!$ 个排列都对应同一个 $\{c_{left}^{'}\}$，所以有：

$$T_m^n = \frac{R_{m-n}}{(m-n)!} = \frac{T_m^m - \sum\limits_{k=0}^{m-n-1} T_{n+k}^n T_{m-n+k}^{m-n}}{(m-n)!} \qquad T_0^0 = 1 \tag{3}$$

使用公式 (1)和公式 (3)能递归地计算任意的 T_m^n。

现在可以计算 $\{a\}$ 与 $\{b\}$ 碰撞 0 次的概率了。因为 T_{n+k}^n 表示的是 $\{a\}$ 与 $\{b^{'}\}$ 碰撞次数为 0 的个数，而 $\{b^{'}\}$ 中有 k 项来自 $\{c_{right}\}$，这 k 项共有 C_{m-n}^k 种组合，因此 $\{a\}$ 与 $\{b^{'}\}$ 碰撞次数为 0 的个数为：$C_{m-n}^k T_{n+k}^n$，对 k 求和可得 $\{a\}$ 与 $\{b\}$ 碰撞 0 次的个数：$\sum\limits_{k=0}^{m-n} C_{m-n}^k T_{n+k}^n$，碰撞 0 次的概率为：

$$p_{n,m,2}(0) = \frac{\sum\limits_{k=0}^{m-n} C_{m-n}^k T_{n+k}^n}{\dfrac{m!}{(m-n)!}} \tag{4}$$

然后计算碰撞 i 次的概率 $(0 \le i \le n)$。当 $\{a\}$ 与 $\{b\}$ 碰撞 i 次时，发生碰撞的 i 项一定是 $\{a\}$ 中的某 i 项，这就有 C_n^i 种可能。

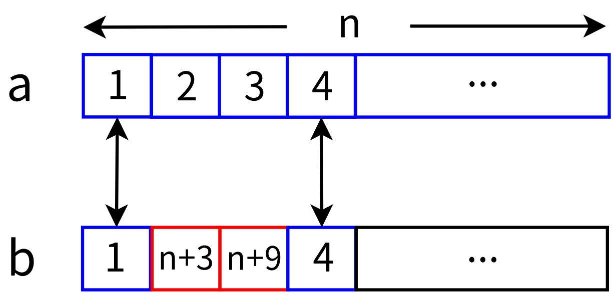

图 4: a 与 b 的碰撞

因为碰撞 i 次，所以 $\{a\}$ 中剩下的 $n-i$ 项与 $\{b\}$ 中剩下的 $n-i$ 项碰撞次数为 0，这就相当于 $n^{'} = n-i, m^{'} = m-i$ 时计算碰撞次数为 0 的个数，因此 $\{a\}$ 与 $\{b\}$ 碰撞 i 次的个数为：
$C_n^i \sum\limits_{k=0}^{m^{'}-n^{'}} C_{m^{'}-n^{'}}^k T_{n^{'}+k}^{n^{'}} = C_n^i \sum\limits_{k=0}^{m-n} C_{m-n}^k T_{n-i+k}^{n-i}$，碰撞 i 次的概率为：

$$p_{n,m,2}(i) = \frac{C_n^i \sum\limits_{k=0}^{m-n} C_{m-n}^k T_{n-i+k}^{n-i}}{\dfrac{m!}{(m-n)!}} \tag{5}$$

问题 2：计算多个序列碰撞的概率

$\{a\}$ 是一个长度为 n 的数列，$a_1, a_2, ..., a_n$ 是 $1, 2, ..., n$ 的某个排列。$\{b^1\}, \{b^2\}, \dots, \{b^t\}$ 也都是长度为 n 的数列，$b_1^j, b_2^j, ..., b_n^j$ 是 $1, 2, ..., n, ..., m$ 的某个排列的前 n 项 $(1 \le j \le t, n \le m)$，($1, 2, ..., n$ 的某个排列和 $1, 2, ..., n, ..., m$ 的某个排列都是从所有排列中独立、等概率选取的)。若存在某个正整数 k 使得 $a_k = b_k^1$ 或 $a_k = b_k^2 \dots$ 或 $a_k = b_k^t$，则称这 $t+1$ 个序列碰撞了一次。求这 $t+1$ 个序列碰撞 i 次的概率 $p_{n,m,t+1}(i)$ $(0 \le i \le n)$。

解

显然 $\{a\}$ 仍然可以是固定的，设 $a_k = k, 1 \le k \le n$。若 $t+1$ 个序列碰撞了 i 次，则发生碰撞的 i 项一定是 $\{a\}$ 中的某 i 项 (有 C_n^i 种可能)，且 a 与 b^j 的碰撞位于这 i 项内，a 与 b^j 的碰撞次数小于等于 i。要计算 $t+1$ 条数列碰撞 i 次的概率，先确定碰撞的是哪 i 项，这样把 a 与 $b^j (1 \le j \le t)$ 的碰撞就限制在这 i 项中，**在此限制下**，先计算 $t+1$ 条数列碰撞次数小于等于 i 次的概率，再从中减去碰撞 $0, 1, \dots, i-1$ 次的概率。最后乘以 C_n^i 得到 $t+1$ 条数列碰撞 i 次的概率。

记碰撞限制在某 i 项中时，$t+1$ 条数列碰撞次数小于等于 i 次的概率为 $q_{n,m,t+1}(i)$，由公式 (5) 可得 a 与 b^j 碰撞次数小于等于 i 次的概率为：$\sum\limits_{z=0}^{i} \dfrac{C_i^z \sum\limits_{k=0}^{m-n} C_{m-n}^k T_{n-z+k}^{n-z}}{\dfrac{m!}{(m-n)!}}$，$(C_0^0 = 1)$。$t+1$ 条数列碰撞次数小于等于 i 次的概率为：

$$q_{n,m,t+1}(i) = \left(\sum_{z=0}^{i} \frac{C_i^z \sum\limits_{k=0}^{m-n} C_{m-n}^k T_{n-z+k}^{n-z}}{\dfrac{m!}{(m-n)!}} \right)^t \tag{6}$$

记 $t+1$ 条数列碰撞次数等于 z，且碰撞在固定的 $z (0 \le z \le n)$ 项中的概率为 $w_{n,m,t+1}(z)$，则有：

$$q_{n,m,t+1}(i) = \sum_{z=0}^{i} C_i^z w_{n,m,t+1}(z)$$

$$w_{n,m,t+1}(i) = q_{n,m,t+1}(i) - \sum_{z=0}^{i-1} C_i^z w_{n,m,t+1}(z),\ w_{n,m,t+1}(0) = q_{n,m,t+1}(0) \tag{7}$$

使用公式 (6)和公式 (7)能递归地计算任意的 $w_{n,m,t+1}(i)$。

最后乘以 C_n^i 得到 $t+1$ 条数列碰撞 i 次的概率：

$$p_{n,m,t+1}(i) = C_n^i w_{n,m,t+1}(i) \tag{8}$$

以上公式的实现和碰撞概率的仿真代码在:
https://github.com/piggypiggy/collision-probability