Asymptotic enumeration of contingency tables

Catherine Greenhill

School of Mathematics and Statistics
University of New South Wales

Joint work with Brendan McKay
(Australian National University)

A contingency table is a non-negative integer matrix with given row and column sums:

$$\begin{array}{cc} \left[\begin{array}{c} \text{nonnegative} \\ \text{integer} \\ \text{entries} \end{array}\right] & \begin{array}{c} s_1 \\ s_2 \\ \vdots \\ s_m \end{array} \\ t_1 \quad t_2 \quad \ldots \quad t_n & \end{array}$$

They are very important in statistics where they are also called frequency tables.

Much interest in efficiently sampling and counting contingency tables. We focus on counting.

Algorithmic results: goal = FPRAS

Constant number of rows:

- Cryan, Dyer, Goldberg, Jerrum and Martin (2002), Markov chain approach.
- Cryan and Dyer (2002), dynamic programming and volume estimation.
- Dyer (2003), dynamic programming algorithm.

"Dense" tables: reduction to volume estimation.

- Dyer, Kannan & Mount (1997)
- Morris (1999)

Asymptotic enumeration.

Let $M(\vec{s},\vec{t})$ be the number of contingency tables with row sums given by $\vec{s} = (s_1, s_2, \ldots, s_m)$ and column sums given by $\vec{t} = (t_1, t_2, \ldots, t_n)$.

Now, rather than an approximation algorithm, we seek a formula for $M(\vec{s},\vec{t})$ with relative error $o(1)$ as $m, n \to \infty$.

Define $s = \max_i s_i$, $\quad t = \max_j t_j$, and

$$S = s_1 + \cdots + s_m = t_1 + \cdots + t_n.$$

History

The matrix is semiregular if $s_i = s$ for all i, $\quad t_j = t$ for all j. In this case write $M(m, s; n, t)$ instead of $M(\vec{s}, \vec{t})$.

Read (1958): asymptotics of $M(n, 3; n, 3)$ as $n \to \infty$.

Everett & Stein (1971), Békéssy, Békéssy & Komlós (1972), Bender (1974): asymptotics of $M(m, s; n, t)$ for bounded s, t.

Now allow s, $t \to \infty$ with m, n. Canfield & McKay (2007+) proved an asymptotic formula for $M(m, s; n, t)$ which holds when the matrices are sufficiently dense. Their proof uses analytic methods.

Canfield & McKay conjectured that $M(m,s;n,t)$ can always be written in a certain form. Conjecture proved for $m=n\leq 9$ using exact values (Beck & Pixton 2003), and computationally for several thousand values of $(m,s;n,t)$ with $m,n\leq 30$.

We verified the conjecture in the case that $st=o((mn)^{1/5})$ (sparse matrices). Further, our asymptotic expression for $M(\vec{s},\vec{t})$ holds for the irregular case when $1\leq st=o(S^{2/3})$.

Our result

Define

$$\mu = \frac{mn}{S(mn+S)} \sum_i (s_i - S/m)^2, \qquad \nu = \frac{mn}{S(mn+S)} \sum_j (t_j - S/n)^2.$$

Suppose that $m, n \to \infty$, $S \to \infty$ and $1 \leq st = o(S^{2/3})$.

If $(1+\mu)(1+\nu) = O(S^{1/3})$ then $M(\vec{s}, \vec{t})$ equals

$$\frac{\prod_{i=1}^{m} \binom{n+s_i-1}{s_i} \prod_{j=1}^{n} \binom{m+t_j-1}{t_j}}{\binom{mn+S-1}{S}} \exp\left(\frac{1}{2}(1-\mu)(1-\nu) + O\left(\frac{st}{S^{2/3}}\right)\right)$$

and otherwise a similar expression holds with the same sized error but with (many) more terms in the $\exp(\cdot)$.

Interpretation: Write $M(\vec{s},\vec{t}) = M P_1 P_2 E$ where

$$M = \binom{mn+S-1}{S},$$

$$P_1 = M^{-1} \prod_{i=1}^{m} \binom{n+s_i-1}{s_i}, \quad P_2 = M^{-1} \prod_{j=1}^{n} \binom{m+t_j-1}{t_j},$$

$$E = \exp\left(\frac{1}{2}(1-\mu)(1-\nu) + O\left(\frac{st}{S^{2/3}}\right)\right).$$

Then M is the number of $m \times n$ nonnegative integer matrices whose entries sum to S. In the uniform probability space on these matrices, P_1 is the probability that the row sums equal $\vec{s}$ and P_2 is the probability that the column sums equal $\vec{t}$.

Hence E is a correction to account for the non-independence of these events!

Let $\mathcal{M}(\vec{s},\vec{t})$ be the set of contingency tables with row sums $\vec{s}$, column sums $\vec{t}$. We need to establish three key facts:

Claim: with probability $1 - O(s^3t^3/S^2)$, a randomly chosen element of $\mathcal{M}(\vec{s},\vec{t})$ has

* no entry greater than 3,

* "not many" entries equal to 3,

* "not many" entries equal to 2.

We prove the key facts using switchings.

Then we borrow calculations from Greenhill, McKay and Wang (2006) (sparse 0-1 matrices) to finish the job.

Key tool

Switchings Theorem (Fack & McKay 2007)
Let $G = (V, E)$ be a finite simple acyclic digraph where each $v \in V$ corresponds to a set $C(v)$, all pairwise disjoint.

Let $\mathcal{S}$ be a set of ordered pairs such that for each $(Q, R) \in \mathcal{S}$ there exists $vw \in E$ with $Q \in C(v)$, $R \in C(w)$.

Also take positive functions $a, b : V \to \mathbb{R}$ such that

$$|\{(Q, R) \in \mathcal{S} \mid Q \in C(v)\}| \geq a(v)\, |C(v)|,$$
$$|\{(Q, R) \in \mathcal{S} \mid R \in C(v)\}| \leq b(v)\, |C(v)|....$$

...Let $\emptyset \neq Y \subseteq V$. Then there exists a directed path $v_1, \ldots, v_k$ in G with $v_1 \in Y$, where v_k is a sink, such that

$$\frac{\sum_{v \in Y} |C(v)|}{\sum_{v \in V} |C(v)|} \leq \frac{\sum_{v_i \in Y} N(v_i)}{\sum_{1 \leq i \leq k} N(v_i)}$$

where

$$N(v_1) = 1, \quad N(v_i) = \frac{a(v_1)a(v_2)\ldots a(v_{i-1})}{b(v_2)b(v_3)\ldots b(v_i)}, \quad 2 \leq i \leq k.$$

Our switchings

For $D \geq 2$ define a D-switching:

$$Q = \begin{pmatrix} D & 0 & 0 & \dots & 0 \\ 0 & q_1 & * & \dots & * \\ 0 & * & q_2 & \dots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \dots & q_D \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & q_1 - 1 & * & \dots & * \\ 1 & * & q_2 - 1 & \dots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & * & * & \dots & q_D - 1 \end{pmatrix} = R$$

Here $q_i \notin \{0, D+1\}$ for $1 \leq i \leq D$. A D-switching preserves all row and column sums as well as the number of entries greater than D. The number of entries equal to D is reduced by at least 1 and at most $D+1$.

The reverse operation is called (wait for it) a reverse D-switching.

For $k \geq 1$ let $S_k = \sum_i [s_i]_k$, $T_k = \sum_j [t_j]_k$.

Lemma. Fix $D \geq 2$ and suppose that $Q \in \mathcal{M}(\vec{s}, \vec{t})$ has at least $K \geq 2st$ positive entries not greater than D and at least J entries equal to D. Then there are at least $J(K - 2st)^D$ D-switchings and at most $S_D T_D$ reverse D-switchings which can be applied to Q.

(This gives the functions $a(v)$, $b(v)$ which we need for Fack & McKay's Switchings Theorem.)

Lemma. Let $\mathcal{U}_1$ be the set of all matrices in $\mathcal{M}(\vec{s}, \vec{t})$ with at least one entry greater than 3. Then

$$\frac{|\mathcal{U}_1|}{M(\vec{s}, \vec{t})} = O\left(\frac{s^3 t^3}{S^2}\right).$$

Proof. Let $\Delta = \min\{s, t\}$ be the largest possible entry. We apply the following argument for $D = \Delta, \Delta - 1, \ldots, 4$ to show that very few matrices have largest entry equal to D.

Define $\mathcal{M}_D(j)$ to be the set of matrices in $\mathcal{M}(\vec{s}, \vec{t})$ with exactly j entries equal to D and none greater. Also let

$$\mathcal{M}_D(>0) = \bigcup_{j>0} \mathcal{M}_D(j).$$

Then $\mathcal{M}_{D+1}(0) = \mathcal{M}_D(0) \cup \mathcal{M}_D(>0)$.

Fix D with $4 \leq D \leq \Delta$. We wish to bound

$$\frac{|\mathcal{M}_D(>0)|}{M(\vec{s},\vec{t})} \leq \frac{|\mathcal{M}_D(>0)|}{|\mathcal{M}_{D+1}(0)|}.$$

Take the digraph with vertex set $V = \{v_0, v_1, \ldots\}$ (where v_j is associated with $C(v_j) = \mathcal{M}_D(j)$) and edge set $E = \{v_j v_i \mid j - D - 1 \leq i \leq j - 1\}$.

Let $\mathcal{S} = \{(Q, R) \mid R$ can be obtained from Q using a D-switching$\}$. Take $Y = \{v_1, v_2, \ldots\} \subseteq V$.

Using the previous lemma, take $a(v_j) = j(S/D - 2st)^D$, $b(v_j) = S_D T_D$ and note $S_D T_D > 0$ since $D \leq \Delta$.

Then the Switchings Theorem says there exists a directed path $v_{t_1}, v_{t_2}, \dots, v_{t_q}$ where $t_1 > t_2 > \cdots > t_q = 0$ (as v_0 is the only sink) and $q > 1$, such that

$$\begin{aligned}
\frac{|\mathcal{M}_D(>0)|}{|\mathcal{M}_{D+1}(0)|} = \frac{\sum_{v\in Y}|C(v)|}{\sum_{v\in V}|C(v)|} &\le \frac{N(v_{t_{q-1}}) + \cdots + N(v_{t_1})}{N(v_{t_q}) + \cdots + N(v_{t_2})} \\
&\le \max_{2\le i\le q} \frac{N(v_{t_{i-1}})}{N(v_{t_i})} \\
&= \max_{2\le i\le q} \frac{b(\mathcal{M}_D(t_i))}{a(\mathcal{M}_D(t_{i-1}))} \\
&= \frac{S_D T_D}{t_{q-1}\,(S/D - 2st)^D} \\
&\le \frac{S_D T_D}{(S/D - 2st)^D} =: \xi_D
\end{aligned}$$

Now

$$\xi_4 = \frac{S_4 T_4}{(S/4 - 2st)^4} \le \frac{s^3 t^3 S^2}{(S/4 - 2st)^4} = O(s^3 t^3 / S^2)$$

and $\xi_{D+1}/\xi_D = o(1)$ for $4 \le D < \Delta$. Hence

$$\frac{|\mathcal{U}_1|}{M(\vec{s}, \vec{t})} \le \sum_{D=4}^{\Delta} \frac{|\mathcal{M}_D(>0)|}{|\mathcal{M}_{D+1}(0)|} \le \sum_{D=4}^{\Delta} \xi_D = O\left(\frac{s^3 t^3}{S^2}\right)$$

as required. $\square$

The argument to show "not many" entries equal to 2 or 3 is a similar but slightly more delicate application of Fack & McKay's Switchings Theorem.

Now we have our nice asymptotic formula. But is any of this practical??

NO! It would be a nightmare to calculate the implicit constant in our $O(\cdot)$ error. So we can't guarantee relative error less than $1+\varepsilon$ for a given ε.

YES! (Kind of) In the semiregular case, if $st = o((mn)^{1/5})$ then our verification of C& M's conjecture strongly suggests relative error at most $\exp(4/(m+n))$. (Here we assume only that a $o(1)$ term is at most 1.)

More DATA needed.